

# NAVAL POSTGRADUATE SCHOOL

## Monterey, California



## THESIS

### DATA WAREHOUSING AND DATA QUALITY FOR A SPATIAL DECISION SUPPORT SYSTEM

by

Robert W. Dill

September 1997

Thesis Advisor:  
Co-Advisor:  
Associate Advisor:

Daniel R. Dolk  
George W. Thomas  
Kathryn Kocher

Approved for Public release; distribution is unlimited.

19980212 054

UNCLASSIFIED//FOR OFFICIAL USE ONLY

# REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE  
September 1997

3. REPORT TYPE AND DATES COVERED  
Master's Thesis

4. TITLE AND SUBTITLE  
**DATA WAREHOUSING AND DATA QUALITY FOR A SPATIAL  
DECISION SUPPORT SYSTEM**

5. FUNDING NUMBERS

6. AUTHOR(S)  
Robert W. Dill

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)  
Naval Postgraduate School  
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION  
REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING  
AGENCY REPORT NUMBER

## 11. SUPPLEMENTARY NOTES

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

## 12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public release; distribution is unlimited.

## 12b. DISTRIBUTION CODE

## 13. ABSTRACT (maximum 200 words)

This research investigates the problems inherent in Decision Support Systems (DSS) that depend on the quality and accuracy of legacy information as the basis for decision-making. A Spatial Decision Support System (SDSS) was developed at Naval Postgraduate School to analyze the comparative desirability of Army Reserve Unit locations. The Army Reserve Installation Evaluation System (ARIES) integrates a GIS mapping engine and a decision model solver in a flexible environment that leverages operational legacy database information for decision-making.

Data quality problems from legacy sources motivated the development of a data migration plan to transform the source data into an architecture optimized for the ARIES SDSS application. This research developed a prototype Data Migration Tool (DMT) to extract the relevant source data into a centralized repository for the SDSS with an acceptable degree of data quality to support SDSS outcomes. Six data quality attributes were identified: accuracy, completeness, consistency, timeliness, uniqueness, and validity. The ARIES DMT focused on data validity and developed techniques for measuring and enforcing data validity. The DMT also specified individual responsibilities for data administration, development of data retrieval routines, and data quality assessment.

Significant system performance enhancements resulted from implementation of the DMT by leveraging the spatial aspects of the underlying repository through geographic queries that efficiently localized subsets of the data files. Additional performance enhancements were obtained through the use of data warehousing techniques.

## 14. SUBJECT TERMS

Data Warehousing, Data Marting, Data Quality, Decision Support Systems, Spatial Decision Support System, Data Migration, Army Reserve, Site Location, Readiness

## 15. NUMBER OF PAGES

226

## 16. PRICE CODE

17. SECURITY  
CLASSIFICATION OF REPORT

Unclassified

18. SECURITY  
CLASSIFICATION OF THIS  
PAGE

Unclassified

19. SECURITY  
CLASSIFICATION OF  
ABSTRACT

Unclassified

20. LIMITATION OF  
ABSTRACT

UL



Approved for Public release; distribution is unlimited.

**DATA WAREHOUSING AND DATA QUALITY  
FOR A SPATIAL DECISION SUPPORT SYSTEM**

Robert W. Dill  
Lieutenant Commander, United States Navy  
B.S., United States Naval Academy, 1985

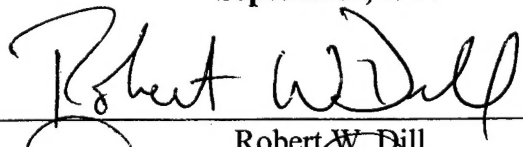
Submitted in partial fulfillment of the  
requirements for the degree of

**MASTERS OF SCIENCE IN  
INFORMATION TECHNOLOGY MANAGEMENT**

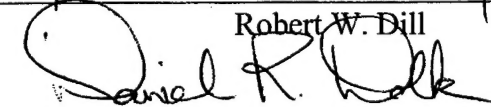
from the

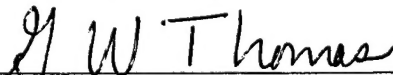
**NAVAL POSTGRADUATE SCHOOL  
September, 1997**

Author:

  
Robert W. Dill

Approved by:

  
Daniel R. Dolk, Thesis Advisor

  
George W. Thomas, Co-Advisor

  
Kathryn Kocher, Associate Advisor

  
Reuben Harris, Chairman,

Department of Systems Management





## ABSTRACT

This research investigates the problems inherent in Decision Support Systems (DSS) that depend on the quality and accuracy of legacy information as the basis for decision-making. A Spatial Decision Support System (SDSS) was developed at Naval Postgraduate School to analyze the comparative desirability of Army Reserve Unit locations. The Army Reserve Installation Evaluation System (ARIES) integrates a GIS mapping engine and a decision model solver in a flexible environment that leverages operational legacy database information for decision-making.

Data quality problems from legacy sources motivated the development of a data migration plan to transform the source data into an architecture optimized for the ARIES SDSS application. This research developed a prototype Data Migration Tool (DMT) to extract the relevant source data into a centralized repository for the SDSS with an acceptable degree of data quality to support SDSS outcomes. Six data quality attributes were identified: accuracy, completeness, consistency, timeliness, uniqueness, and validity. The ARIES DMT focused on data validity and developed techniques for measuring and enforcing data validity. The DMT also specified individual responsibilities for data administration, development of data retrieval routines, and data quality assessment.

Significant system performance enhancements resulted from implementation of the DMT by leveraging the spatial aspects of the underlying repository through geographic queries that efficiently localized subsets of the data files. Additional performance enhancements were obtained through the use of data warehousing techniques.



## TABLE OF CONTENTS

<b>I. INTRODUCTION .....</b>	<b>1</b>
A. GENERAL.....	1
B. BACKGROUND .....	2
C. THESIS OBJECTIVES .....	3
D. SCOPE.....	3
E. ORGANIZATION OF THE STUDY .....	4
<b>II. SDSS ARCHITECTURE DEVELOPMENT .....</b>	<b>5</b>
A. THE ARIES PROJECT .....	5
B. DECISION MODEL.....	6
1. Decision Process Elicitation .....	6
2. Decision Goals and Goal Hierarchy .....	7
3. Decision Measures.....	10
C. DATA MODEL .....	11
1. Developing the Business Rules.....	11
2. Mapping the Business Rules to Real Data .....	12
3. Identify Source Data .....	12
D. SYSTEM ARCHITECTURE .....	13
1. Integrating Shell.....	14
2. Mapping Engine .....	16
3. Decision Model Solver .....	16
4. Data Preprocessor .....	17
E. DATA PREPROCESSOR: ARIES ADMINISTRATOR .....	18
1. Maintaining File Locations.....	18
2. Query Development Process.....	18
3. Data Extraction Queries.....	19
4. Data Cleaning, Standardizing, and Extracting .....	20
F. CHAPTER SUMMARY .....	21
<b>III. DATA WAREHOUSING AND DATA QUALITY: TOPICAL DISCUSSION .....</b>	<b>23</b>
A. DATA WAREHOUSING.....	23
1. Definition.....	23
2. Applications.....	26
3. Design Concerns .....	28
B. DATA MARTS.....	29

1. Definition.....	30
2. Applications.....	32
3. Design Concerns.....	33
C. DATA QUALITY.....	34
1. Definition.....	35
2. Improving Data Quality.....	37
3. Data Migration.....	38
D. CHAPTER SUMMARY .....	39
<b>IV. ARIES SDSS APPLICATION: PROBLEM EVALUTION .....</b>	<b>41</b>
A. BUSINESS RULE DEVELOPMENT PROBLEMS.....	41
1. Errors in Logic.....	42
2. Rule not Supported by Data.....	43
B. QUERY PERFORMANCE PROBLEMS .....	43
1. SQL vs. Geo-Query .....	44
2. Detail vs. Aggregation .....	45
C. DATA ANOMALIES.....	46
1. Proxy Value Calculations .....	46
2. Data Validation.....	47
3. Data Quality Analysis.....	52
D. CHAPTER SUMMARY .....	54
<b>V. LESSONS LEARNED: SDSS DESIGN AND DEVELOPMENT .....</b>	<b>57</b>
A. SDSS DEVELOPMENT PROCESS .....	57
B. DATA MIGRATION PLAN .....	59
1. Designate Migration Team .....	59
2. Determine Extraction Logic and Generate Extraction Routine .....	60
3. Quality Assured Data.....	61
C. DECISION MODEL DEVELOPMENT .....	62
D. DATA MODEL .....	63
1. Data Standardization.....	63
2. Meta-data Documentation Process .....	64
3. Identify Spatial Aspects of Queries .....	65
E. SYSTEM DESIGN .....	66
F. FUTURE CONSIDERATIONS .....	66
1. Decision Model.....	66

2. Data Migration .....	67
3. Data Model .....	67
4. System Design .....	68
5. Testing .....	69
G. CHAPTER SUMMARY .....	69
<b>VI. CONCLUSION .....</b>	<b>73</b>
A. SUMMARY .....	73
B. CONTRIBUTIONS .....	74
1. General Contributions .....	74
2. Specific Contributions to USARC .....	75
<b>APPENDIX A. DECISION MODEL MEASURES .....</b>	<b>77</b>
<b>APPENDIX B. ARIES SOURCE DATA FILE META-DATA .....</b>	<b>121</b>
<b>APPENDIX C. ARIES DECISION MEASURE STATISTICS .....</b>	<b>157</b>
<b>APPENDIX D. SOURCE FILE DOCUMENTATION FORMS .....</b>	<b>199</b>
<b>LIST OF REFERENCES .....</b>	<b>203</b>
<b>BIBLIOGRAPHY .....</b>	<b>205</b>
<b>INITIAL DISTRIBUTION LIST .....</b>	<b>207</b>



## LIST OF FIGURES

Figure 1. ARIES Architecture Development Process .....	5
Figure 2. ARIES System Architecture.....	14
Figure 3. ARIES User Interface Screen for Specifying Parameters .....	15
Figure 4. ARIES Administrator File Location Screen.....	19
Figure 5. ARIES Administrator Extract Queries Screen .....	20
Figure 6. Example Fact Table. [Ref. 9:Figure 2].....	27
Figure 7. Data Migration Process .....	38
Figure 8. Recommended SDSS Development Process.....	58



U

## LIST OF TABLES

Table 1. Complete Hierarchy of Goals for Site Desirability.....	8
Table 2. Goal Hierarchy (showing only those measures with automated inputs).....	9
Table 3. Data Warehouse, Data Mart differences. [Ref. 10:p. 9].....	30
Table 4. Data Quality Characteristics. [Ref. 16:p. 2].....	36
Table 5. VALID UIC Query .....	42
Table 6. COMMAND PLAN Filter Query .....	43
Table 7. Example Aggregation Query - G18NatI UIC .....	45
Table 8. Descriptive Statistics for Decision Measures .....	48
Table 9. ARIES Measures Analysis Statistics - Run #1 .....	49
Table 10. ARIES Measures Analysis Statistics - Run #2 .....	51



## LIST OF ACRONYMS

ADM	Advanced Development Model
AMSA	Area Maintenance Support Activity
ARIES	Army Reserve Installation Evaluation System
COTS	Commercial Of-The-Shelf
DISA	Defense Information Systems Agency
DMP	Data Migration Plan
DSS	Decision Support System
ECS	Equipment Concentration Center
EDM	Enterprise Data Model
FACID	Facility Identification Code
FSP	Force Support Package
GIS	Geographical Information System
GUI	Graphical User Interface
IRR	Individual Ready Reserve
LAN	Local Area Network
LDW	Logical Decision for Windows
MDDB	Multi-Dimensional Database
MOS	Military Occupational Specialties
NPS	Naval Postgraduate School
OLAP	On-line Analytical Processing
OLTP	On-line Transactional Processing
OLE	Object Linking and Embedding
RDBMS	Relational Database Management System
ROLAP	Relational On-line Analytical Processing
SDSS	Spatial Decision Support System
SQL	Structured Query Language
TDQM	Total Data Quality Management
TPU	Troop Program Unit
UDA	Unified Data Architecture
UIC	Unit Identification Code

USARC

United States Army Reserve Command

## ACKNOWLEDGEMENT

First and foremost I want to thank my family for supporting me in continuing my education. Without their faithful support and love, I know I would not be where I am today. Thank you Michele, R.J., and Patrick for allowing me the time and supporting me in this project. I also want to thank my Mom and Dad for stressing the importance of education in my life.

I would like to thank Professors Dolk and Thomas for giving me the opportunity to work on this project and for allowing me to take the topic of my thesis in the direction that I wanted. The knowledge that I have gained in conjunction with this project can not be documented in a paper.

Thank you to Kathryn Kocher for her assistance as an additional reader. Her insights and guidance were very important to the final product of this thesis. She offered a valued opinion of someone outside the details of the ARIES project.

The development of the ARIES prototype application may not have been possible without the generosity of Professor Barry Frew. He allowed the development team full use of a desktop computer that was essential to the success of the overall project.

Professor Shu Liao also provided a valuable resource, in the form of a laptop computer, that the project and my thesis could not have been completed without. Because of his generosity we were able to run the application on the desktop for long periods of time and conduct analysis from the laptop.



# **I. INTRODUCTION**

## **A. GENERAL**

This research analyzes the problems inherent in Decision Support Systems (DSS) that rely upon legacy databases as the primary data source. The quality and accuracy of an outcome that any DSS returns cannot be better than the quality and accuracy of the underlying database information. We explore this premise in the context of a prototype Spatial Decision Support System (SDSS) developed for the United State Army Reserve Command (USARC) that allows analysis of the comparative desirability of Army Reserve Unit locations. Since many DSS's are model-based, initial development often focuses on specification of the underlying model(s) and associated user interfaces. Issues concerning the data required to run the models are frequently left until the latter stages of development.

Initial implementation of the USARC SDSS took exactly this approach and, as a result, encountered serious problems with the underlying data that compromised the quality of the decisions that the SDSS was able to render. Significant measures were required to resolve these problems; specifically an entire data administration module had to be developed to identify meta-data and regulate the extraction of DSS data from source data files. This module required the adoption of procedures to assess and monitor the quality of the data as it passed from the source legacy databases to the databases used as input to the SDSS. The spatial nature of much of the data put a special twist into this process since the SDSS application takes advantage of these spatial aspects to streamline system performance. This research explores the confluence of data quality, decision support, legacy data, and spatial data, and prescribes procedures for dealing with data quality in SDSS development. A major lesson learned was data quality must be addressed at the beginning of (S)DSS projects and not left until the end of the development cycle.



## **B. BACKGROUND**

The Force Support Package (FSP) Readiness Office, a component of the U.S. Army Reserve Command (USARC), is tasked with assessing and improving the readiness of priority Troop Program Units (TPU). A TPU is the foundation of the Army Reserve force, ranging from 50 to 250, typically consisting of about 150 reservists. The TPUs that are in the FSP, which contain the units designated for rapid deployment, are of most concern to the Readiness Office.

Readiness, in this context, refers primarily to personnel readiness, i.e., the ability to maintain troops that are properly trained and qualified individuals in a sufficient number. Many of the numerous factors that affect readiness are dependent on the location of the unit. Relocation of a unit to another facility can, at times, be the best solution when a unit is struggling to maintain personnel readiness. During today's environment of force reductions and realignments, relocation may also be necessary to support force consolidation or restructuring efforts.

Previously these decisions were based upon a combination of personal expertise and narrowly focused studies. This ad hoc process produced results that often proved difficult to communicate, defend, and build consensus around. The human decision-maker becomes overloaded quickly by the large number of factors involved in the TPU relocation decision without the aid of an automated decision tool. The inadequacies of the current approach to provide any detailed solutions to such a complicated problem inspired the search for a convenient and systematic automated tool that could be based upon a decision model.

The use of computer based DSS's to aid the decision-maker in making thorough and informed decisions will become more prevalent as the use of distributed working environments increase. Distributed environments allow access to more information that will increase the overall effectiveness of decision support tools. Experience with the USARC SDSS indicates that significant attention must be paid to the quality of data underlying decision systems in order to ensure the quality of the resulting decisions.

### **C. THESIS OBJECTIVES**

The primary objective of this thesis is to identify problems with developing an SDSS based upon legacy databases with a high variance in data quality. A secondary objective is to develop an application design process that, by incorporating data warehousing techniques, can counteract the effects of poor data quality on the resulting application. This involves analyzing the development process used for the current prototype, identifying the relevant data quality factors, reviewing data warehousing techniques, applying those techniques to address data quality problems in the prototype application, and examining lessons learned from the prototype development process. The research questions that will be addressed are:

- What inherent problems are involved in the use of legacy database information in the development of a state of the art DSS?
- What are the relevant data quality factors for site location decision problems?
- What data warehousing techniques are relevant to the SDSS design process?
- What steps should be taken during the design and development process to ensure that the data quality will support a level of confidence required by the user in the outcome decision?
- Who should be responsible for the level of data quality involved in the development of an SDSS?
- What are the unique problems that spatially enabled data present to the level of data quality?

### **D. SCOPE**

This study will focus on the SDSS developed for USARC to support the unit location decision problem and unit readiness mission responsibilities. The automated decision tool supports the process of relocating units that are not meeting readiness goals to sites that afford them better opportunities to succeed.

The USARC prototype, the Army Reserve Installation Evaluation System (ARIES), has a number of external restrictions imposed that limit the true effectiveness of the system. For example, only those facilities currently owned by the Army Reserve are

considered as potential relocation sites (approximately 1,500 nationwide). The discussion of data quality will be in reference to the data that supports the decision factors of these facilities. For further details of the ARIES project refer to references 1 and 2.

The original project requirements intended to avoid any extensions to existing data maintenance responsibilities. USARC also specified that all model inputs would be drawn from existing data sources. ARIES provides the decision-maker the ability to manually input data needed to support additional decision criteria for incorporation in the evaluation process. This off-line analysis would inject even more concern for data quality and the subsequent confidence level of subsequent decisions that can be reached using this tool. This research does not address data quality issues that arise from this source of "ad hoc" data; rather it focuses on the data quality of the "feeder" legacy databases and their extracted counterparts in the SDSS.

The research sponsor did not define adequately the ownership associated with the data that was used to support the majority of the decision criteria. The lack of a responsible custodian left the interpretation of many of the data fields up to the designers and their ability to ascertain the meaning of the underlying database schema.

## **E. ORGANIZATION OF THE STUDY**

The balance of this study is organized as described below. Chapter II discusses the design process and architecture used in the development of the ARIES prototype project. Chapter III discusses the basic characteristics and elements involved in data warehouses, data marts and issues with data quality. Chapter IV details how these data warehousing techniques were implemented in the ARIES SDSS application. Also discussed in that chapter are the problems with the availability and quality of the database information used in the decision process that surfaced throughout the production of the prototype user interface. Chapter V provides a number of post application design issues and recommendations for further study that could assist future SDSSs of this type. Chapter VI presents conclusions and the contributions of this study.

## II. SDSS ARCHITECTURE DEVELOPMENT

### A. THE ARIES PROJECT

This chapter describes the design process and architecture of the ARIES SDSS project developed at the Naval Postgraduate School (NPS) for the Army Reserve Command. The ARIES Development Process is depicted below in Figure 1. At the heart of the ARIES architecture is a decision model that was developed in conjunction with a group of experts at the FSP office. The decision model produces a list of decision criteria that must then be mapped to operational source databases. This mapping process was done by identifying business rules for each criterion in the form of queries. These business rules became the basis of the data model used in the application. Once the

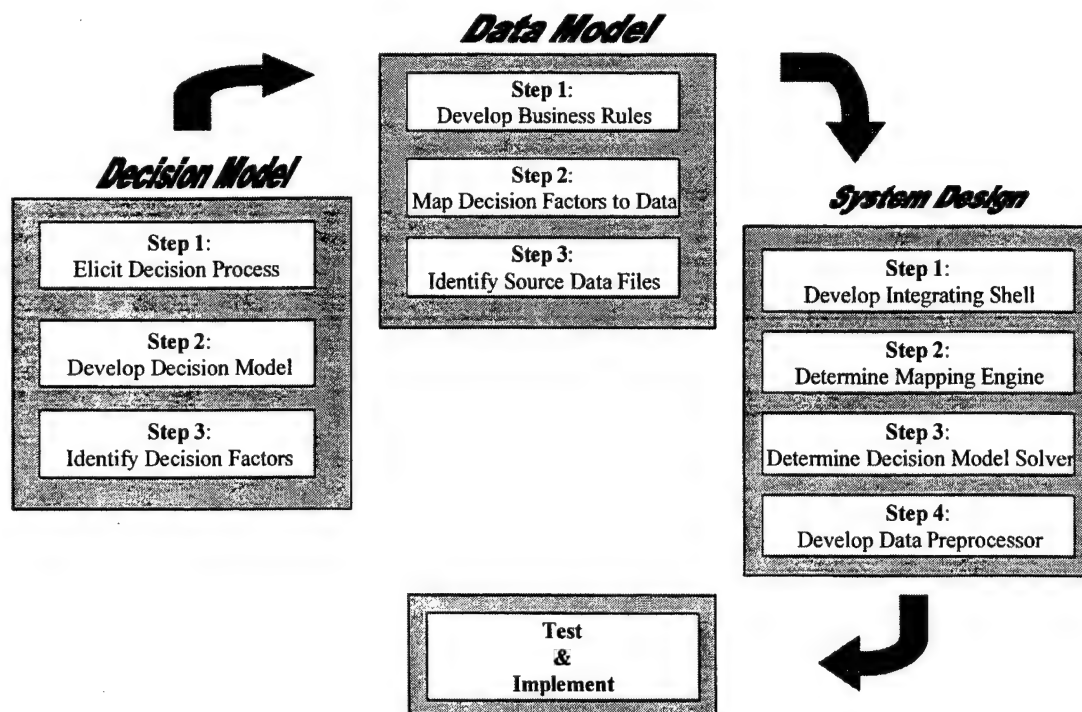


Figure 1. ARIES Architecture Development Process

decision model and data model were finalized, development of the system and user interface began.

## **B. DECISION MODEL**

The core of any DSS is one or more decision models. The ARIES decision model is a multi-criteria model represented as a hierarchy of *objectives* or *goals* with associated *measures* or *criteria* involved in making a specific decision. An *objective* is referred to in most decision literature as a desired direction and a *goal* as the quantifiable progress in that direction. For the purposes of this discussion, we will adopt the terminology of *goals* and *measures* that is consistent with the decision software package used in the ARIES project.

To begin the design of a decision model, a detailed elicitation process is required to capture the characteristics and aspects associated with the specific decision problem being modeled. This process identifies a top-level decision goal that is subsequently refined to layers of subordinate goals. The subordinate goals and their associated decision measures must be arranged in a hierarchy that allows the final analysis to arrive at an evaluation for the top-level goal.

### **1. Decision Process Elicitation**

The first step in modeling the TPU location decision problem was to gather a group of knowledgeable experts in the area of Army Reserve manpower and identify the top-level, or overall goal. The use of experts rather than an extensive study was adopted in the interest of cost savings as well as the ability to develop a working prototype decision model in a short period of time. The elicitation process was done by focusing on factors that were identified in prior research, Reference 1, of the TPU readiness issue as well as the process knowledge of the experts.

The expert panel, consisting of USARC personnel, was able to identify an overall goal that related unit location to unit readiness. This was a challenging process because

of the difficulty of placing a measurement on readiness. Eventually the expert panel settled on an overall goal of site desirability. The panel decomposed site desirability into two subgoals, personnel readiness support (the ability to maintain the desired number of qualified reservists at the proposed site) and facility quality (a general assessment of the costs and benefits of a location that are only loosely related to readiness).

This approach to determining decision goals was initially done without any concern for the availability of data that would subsequently support the model. The elicitation process also did not involve any formal review of the current process for making this decision. Rather, it was an effort to determine the ideal decision process for TPU readiness in the context of unit location. A review of the existing decision process would have identified information currently used to make an informed decision which could later become the foundation for building a data model. The lack of such a detailed data model proved to be an obstacle to the project.

## **2. Decision Goals and Goal Hierarchy**

The overall decision goal of Site Desirability was broken down into two subgoals, Facility Quality and Personnel Readiness, which were in turn refined further into either additional subgoals or decision measures. Decision measures are the basic elements of the model to which a single objective value can be assigned. Each subgoal must be ultimately broken down into these basic elements to allow the multi-criteria decision making to occur. Table 1 shows the breakdown of the facility quality and personnel readiness decision goals into decision measures for the ideal decision model. Subsequent discussions revealed that that data did not exist for some of the measures so they were dropped from the initial model. Table 2 shows the decision measures could to be implemented with available data.

The *Facility Quality* subgoal is used to describe specific attributes of a proposed facility (i.e., the building and the real estate). These values are primarily extracted from

## SITE DESIRABILITY

### *I. Facility Quality*

*% Administrative Space FT*

*% Administrative Space PT*

*% Motorpool Space*

*Distance to Headquarters*

*Facility Age*

*Facility Maintenance Backlog*

*Facility Condition*

*Facility Operating Costs*

*Facility Leased/Owned*

### *II. Personnel Readiness*

*MOS Qualification*

*Available Prior Service*

*Available MOS from Closing Units*

*Available MOS from the IRR*

*Fill Level*

*Market Supportability*

*Market Quality*

*Civilian Labor Market*

*Closing Unit Transfers*

*IRR Availability*

*Recruit Market Size*

*Area Units*

*Area Drill Attendance*

*Area Loss Rate*

*Area Transfer Rate*

*Average Area Manning*

*Distance to Recruiter*

*Reassignments*

*Competition*

*Training Support*

*Equipment Readiness*

*% Storage On-Site*

*Distance to AMSA*

*Distance to ECS*

*Training Facility*

*% Facility Usage*

*Distance to Special Training*

*Distance to WET Site*

*Distance to Weapons Range*

*Facility Weekend Use*

**Table 1. Complete Hierarchy of Goals for Site Desirability**

## SITE DESIRABILITY

### *I. Facility Quality*

*Facility Age*

*Facility Maintenance Backlog*

*Facility Condition*

*Facility Operating Costs*

*Facility Leased/Owned*

### *II. Personnel Readiness*

*MOS Qualification*

*Available Prior Service*

*Available MOS from Closing Units*

*Available MOS from the IRR*

*Fill Level*

*Market Supportability*

*Market Quality*

*Closing Unit Transfers*

*IRR Availability*

*Recruit Market Size*

*Area Units*

*Area Drill Attendance*

*Area Loss Rate*

*Area Transfer Rate*

*Average Area Manning*

*Distance to Recruiter*

*Reassignments*

*Competition*

*Training Support*

*Equipment Readiness*

*Distance to AMSA*

*Distance to ECS*

*Training Facility*

*Facility Weekend Use*

**Table 2. Goal Hierarchy (showing only those measures with automated inputs)**

databases maintained by the Army's Corps of Engineers and describe the age, condition, capacity, and costs associated with the major structures of the site.



The *Personnel Readiness* subgoal is used to determine the ability of the area to support personnel readiness. Personnel readiness was broken down into two subgoals, *Fill Level* and *Military Occupational Specialties (MOS) Qualification Level*. *Fill Level* indicates the ability of the area surrounding a site to support a sufficient number of reservists whereas *MOS Qualification Level* indicates the availability of the skill set required by the moving unit in the area of the proposed site. Each of these goals is further broken down generating a hierarchy of the goals and measures that make up the actual decision model.

The resulting hierarchy of goals represents the location-related factors that were determined by consensus of the expert panel to be important in the TPU relocation decision. This goal hierarchy, shown in Tables 1 and 2, is used by the multi-criteria decision solver to obtain a final evaluation of the desirability of each site.

### **3. Decision Measures**

A decision measure is the result decomposing of each goal in the hierarchy into objective inputs that can be qualified and assessed. These objective inputs can come from various sources such as databases, spreadsheets, data analysis, etc. The hierarchy developed by the expert panel allowed most of the inputs to come from existing database information, minimizing the involvement of the user.

The decision analysis software integrates all the dissimilar dimensions of the measures by obtaining a common unit value for each decision measure. The common unit value is arrived at through the use of yield curves for the decision measures. A relative weight is also applied to each goal to denote the level of importance of that goal compared to other goals. As a result, certain nodes in the hierarchy can be calibrated to affect the outcome more strongly than others by assigning them higher weights. These values are then summed for each goal to determine the overall desirability.

Appendix A gives a detailed summary of each measure including the definition, source information used, resulting queries, and associated yield curve. These decision measures became the foundation that for the application's data.

## **C. DATA MODEL**

*"Good decision support requires an integrated, stable, well-managed data resource."* [Ref. 3:p. 267]

A DSS requires data sources from which to draw information that will fully support the underlying decision model. For the ARIES decision hierarchy to provide acceptable confidence levels for the resulting decisions, it must be based upon objective historical data. USARC stipulated that the ARIES application should minimize any associated data management. Specifically they required that ARIES give rise to no new data administration responsibilities. Further, they specified that all database information used must be available or easily transferable to the USARC Local Area Network (LAN), and the application should be able to retrieve information from those available data sources without regard for their location. These basic requirements formed the foundation from which the ARIES data model was developed.

Using the goal hierarchy and the resulting decision measures, steps were taken to identify "business rules" for each decision measure that could be translated into objective equations. These business rules were also derived by consensus of the expert panel. The business rules allow source data elements to be identified that provide an objective assessment of a measure and therefore automate the site evaluation process. Once the required data elements are identified, the set of data files required to support the application are also identified.

### **1. Developing the Business Rules**

Using the ideal decision hierarchy and resulting decision measures, the expert panel documented the factors or elements comprising each measure (i.e., Average Area

Manning = Number of Personnel Assigned / Number of Personnel Required). In addition to developing an objective rule for each measure, it was necessary for the expert panel to define each element that made up this business rule. These definitions are used to determine the actual data that are required to support each decision measure. A complete description of each measure and the associated business rule is contained in Appendix A.

## **2. Mapping the Business Rules to Real Data**

Each equation definition identified required individual data elements. These data elements were then mapped to operational data elements available in database files on the USARC LAN. It was determined that ten of the decision measures in the ideal decision hierarchy did not have readily available data that could support the model. These measures were not automated in the final prototype application.

Logic diagrams were drafted for each measure using the business rule and the identifying source data files. This information was gathered through discussions with individuals at USARC headquarters familiar with the requisite database information. Who were able to identify specific files that would contain the required data elements for each measure. Some information such as census information and facility information were determined to be available through other sources such as the Corps of Engineers.

Because of the geographic nature of some of the decision measures (e.g., Area Loss Rate), it became apparent that a spatial dimension would be necessary in the final application. This requirement for area-specific data led to the integration of a Geographic Information System (GIS) to aid in the selection, querying, and visualization of this decision problem.

## **3. Identify Source Data**

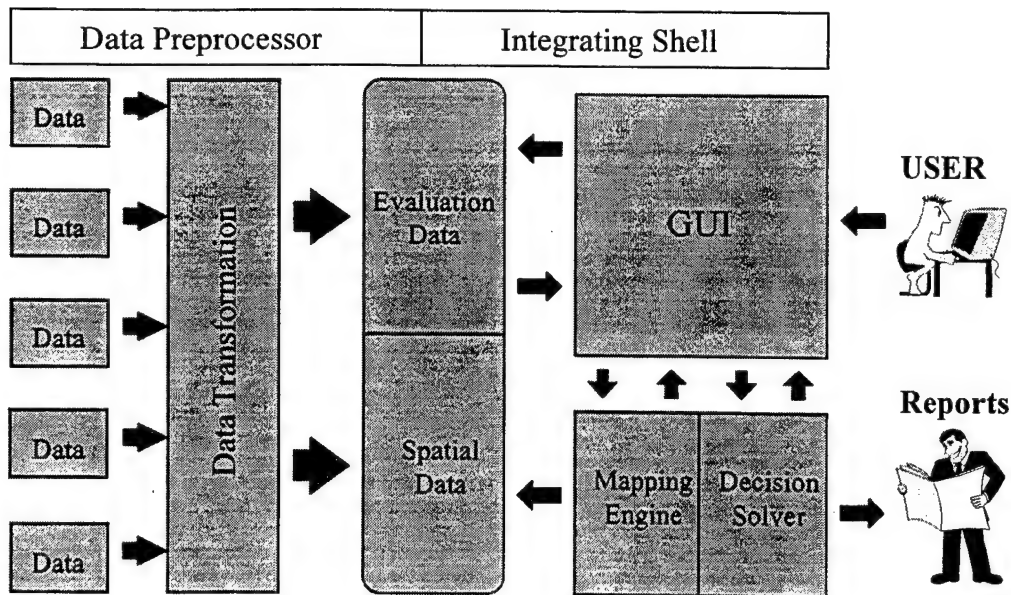
The source data identified include several types of database files, transactional data, spatial data, personnel data, historical data, and analysis data. This wide spectrum of

dissimilar database types posed a challenge in refining the logic developed for each measure. In most cases these source databases were being used and maintained by different entities within the USARC headquarters facility for their own use. The initial data sets that were collected for the prototype were extracts of these databases for the state of Pennsylvania. As development progressed the full national databases were collected and integrated into the project. A description of all the source databases is contained in Appendix B with the meta-data information available for each file. The initial development plan was to draw information directly from each source file during each individual site evaluation session. This process was found to be inefficient as discussed later in the chapter.

#### **D. SYSTEM ARCHITECTURE**

With the Decision Model complete and the Data Model specified, an automated decision application could then be developed to integrate the decision model's information needs with the knowledge of available information in the data model. The Graphical User Interface (GUI) accepts the required inputs for the problem from the user and conducts the evaluation of the defined scenario. This basic architecture is depicted in Figure 2.

It became clear early in the project that budget and time limitations would not allow the development of an application that would carry out all functions of this project independently. This led to the integration of several commercial off-the-shelf (COTS) products to conduct the decision analysis and assist in the GIS portion of the project. The ARIES application architecture consists of four components: an integrating shell, a mapping engine, a decision model solver, and a data preprocessor.



**Figure 2. ARIES System Architecture**

### **1. Integrating Shell**

The application shell that integrates and operates the GUI is original code written in Visual Basic™. Visual Basic™ is an event-driven programming language that allowed the integrating shell to be developed to integrate into the infrastructure already in place at USARC. Visual Basic™ was chosen as the programming language because the USARC information system support personnel were maintaining other applications with it and already had a basic level of understanding. This would allow for future maintenance and improvements to the prototype to be completed in house by USARC personnel.

Another USARC requirement for the prototype was that the final application should relieve the user of the burden of understanding the individual COTS applications and protocols involved in the transfer of information. Because of the predictable and structured nature of this decision process automation of most of the tasks was very effective. The only required inputs from the user are the moving unit identification code (UIC) and the facility identification code (FACID) for the proposed sites. Figure 3 shows the ARIES User Interface screen used to capture the input parameters.

ARIES: v3.0.2  
File LDW View Reports Help

Geoselect Geoquery Measures

**PROPOSED FACILITY INFO**

FAC ID: VA045 UIC: -NONE-

UNITNAME: NORFOLK AFRC

CITY: NORFOLK STATE: VA ZIP: 23511

**ACCEPT**

☐ USARC GeoREF  
☐ VA045  
☐ None

**ARIES Controls**  
Aries Clear Exit

**FACILITY SELECTION**

Moving Unit	Facility One	Facility Two	Facility Three	Facility Four
UIC: W7TCAA	FacID: VA027	FacID: VA022	FacID: VA009	FacID: VA045
FacID: VA032	UIC: -NONE-	UIC: W8XJ20	UIC: W8WR99	UIC: -NONE-
0000 USA ELE HQ AT LANTIC CMD	FORT MONROE USAF	8830 USA (OSUT) BC E 1ST BN HHC	0317 IN RGT 01 (BC T) 28DE AUG	NORFOLK AFRC
FORT STORY VA 23459	FORT MONROE VA 23651	FT EUSTIS VA 23604	SUFFOLK VA 23435	NORFOLK VA 23511

Select ARIES button to start Facility Comparisons Zoom: 70.7 mi Custom Tool: FacID Info

**Figure 3. ARIES User Interface Screen for Specifying Parameters**

The GUI will accept these inputs either as manual inputs or from the map display. The overarching shell uses a set of predefined tasks based on the decision model to acquire the database information for each decision measure. Some of these tasks are carried out through an Objected Linking and Embedding (OLE) connection with the mapping engine and others are carried out using the database engine in Visual Basic™.

Once the shell has obtained values for all the decision measures associated with each proposed site, this information matrix is passed to the decision solver. The decision solver carries out its evaluation and passes control back to the GUI where the user has the ability to print reports, conduct dynamic analysis, or consider another scenario.

## **2. Mapping Engine**

MapInfo™, already in use at USARC, was chosen as the mapping engine for several reasons. MapInfo™ satisfied all the known and anticipated functional requirements, it was already owned by USARC, had proven to be well supported and documented, and would minimize the need for additional training.

MapInfo™ is a commercial mapping package that is used as a graphical input tool and provides for the spatial definition and processing of data. It converts positions to distances, makes proximity determinations, and classifies objects by geographical region. The integrating shell uses the OLE connection to pass data to and from MapInfo™ and launch a MapBasic™ program that executes the spatial queries. The ability of MapInfo™ to localize data from huge databases provided a significant performance gain when the spatial queries were implemented.

## **3. Decision Model Solver**

A decision solver was required that would conduct multi-attribute utility analysis and allow for “what-if” dynamic analysis functionality. Logical Decision for Windows™ (LDW) is used as the decision solver in the ARIES application. LDW™ was chosen primarily for its superior implementation of the underlying decision framework, Multi-Attribute Utility Theory, and its ability to provide a flexible decision analysis environment.

LDW™ was determined to be superior than other similar products in terms of overall ease of use for the novice user. LDW™ supports for a wide range of techniques to obtain user preferences (e.g., ordinal criteria ranking, tradeoffs, direct graphical and tabular inputs). The application also allows the user to set the specific information about the yield curve that affects each decision measure to include: slope, continuous or discrete, minimum and maximum values, and shape. Another important feature that

LDW<sup>TM</sup>, particularly for the ARIES prototype, is the ability to conduct dynamic sensitivity analysis of an evaluation session.

Given these fundamental strengths, LDW<sup>TM</sup> did have limitations in its ability to communicate with the other applications. The ARIES application must pass control to LDW<sup>TM</sup> when the decision analysis phase begins to allow LDW<sup>TM</sup> to work. The 16-bit architecture of LDW<sup>TM</sup> limited the available control methods allowing key-stroke passing as the only means to control the program externally. This limitation requires that the user be familiar with and be able to carry out some functions within LDW<sup>TM</sup> in order to take full advantage of the capabilities of the decision model solver. Through the use of these methods of passing control and information the basic evaluation of a single site location problem is fully automated to include report output.

The ARIES shell passes the subjective values for each decision measure to LDW<sup>TM</sup> for evaluation against the stored default preference set of the goal hierarchy. This is done through a text file because of limitations in LDW<sup>TM</sup>'s capacity to interact with other applications. LDW<sup>TM</sup> receives the matrix of values with the facility name and, using the stored yield curves and assigned weighting, evaluates the specific scenario. The user can either print the standard reports or carry out further analysis of that scenario using the LDW<sup>TM</sup> application.

#### **4. Data Preprocessor**

The final component of the system application, the data preprocessor, evolved from the need to have the operational data move smoothly into the ARIES evaluation process. The data preprocessor, like the shell, is written in Visual Basic<sup>TM</sup>. Even if all source databases were consistent and accurate, their number and sizes present considerable performance challenges for a PC-based, front-end processor. Because of the size and the varying location of the data files involved in the ARIES data model an application that would provide an administrative function for the source information was necessary.



## **E. DATA PREPROCESSOR: ARIES ADMINISTRATOR**

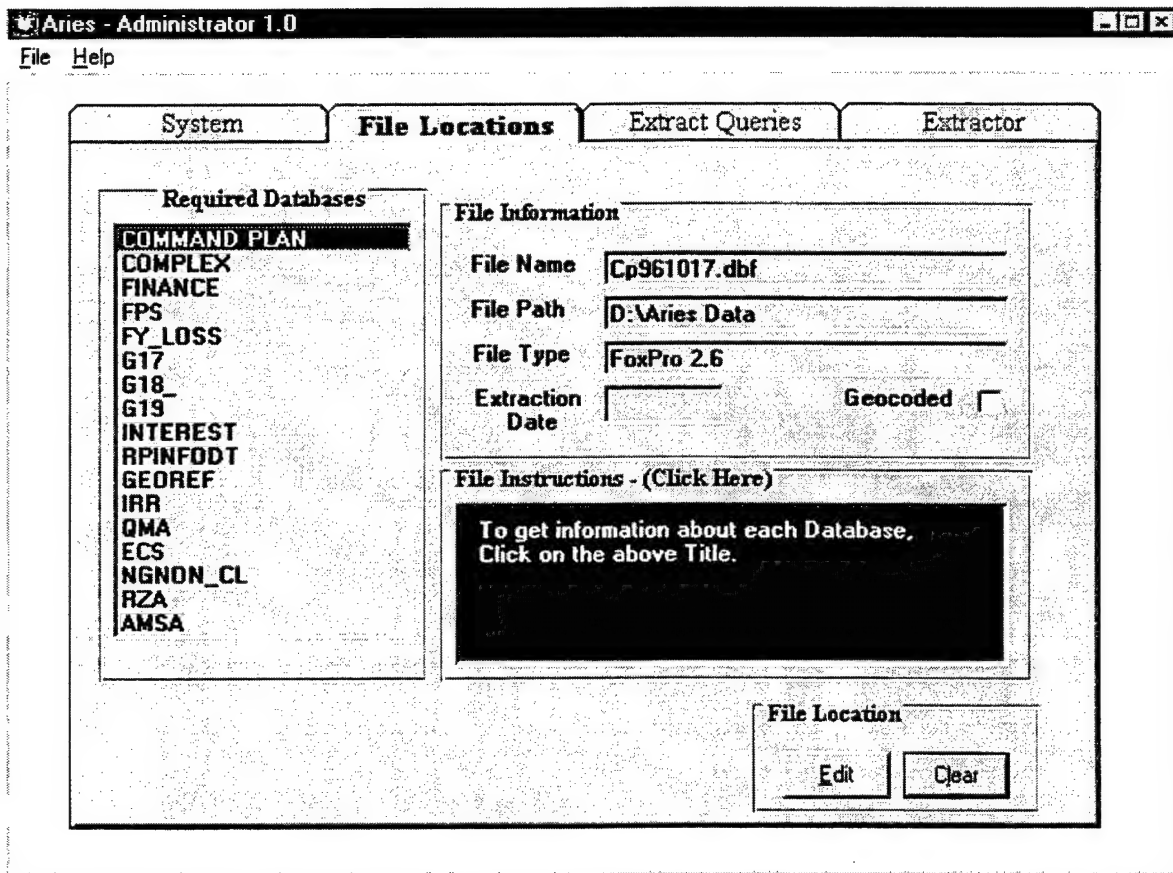
The data preprocessor, known as ARIES Administrator, is the transition element that moves the operational data from its source form to a centralized data resource that the ARIES application can access. USARC's initial requirement to maintain the current location of each source data file was the primary reason for developing this component. As the prototype development progressed, it became clear that all the data elements had to be assembled in one central location to facilitate an acceptable performance level during problem evaluation. This additional function was taken on by the Administrator which evolved into an extracting agent. For the Administrator to conduct an extraction of the source data files, queries had to be generated and maintained in order for the process to be duplicated as the data files changed. The ARIES shell and the Administrator are separate applications that are only connected by the requirements of the ARIES data resource file structure.

### **1. Maintaining File Locations**

In order for the Administrator to find a file for the extraction process the file name, path, and type must be maintained. This information is entered in the Administrator under the File Location tab by using the standard windows file location interface. Figure 4 shows the information maintained for each database. The Administrator also provides, for informational purposes, a list of fields, table names, and table indices that must be present to support the processing performed by the ARIES shell. In addition, the Administrator also maintains a file location of the COTS applications, MapInfo™ and LDW™, to allow flexibility in the installation of these supporting applications.

### **2. Query Development Process**

Development of extract queries became necessary to obtain an acceptable level of performance for the ARIES application. The initial extract queries were designed to



**Figure 4. ARIES Administrator File Location Screen**

retrieve only the required fields and records and place that information in the Microsoft Access™ format. This would allow the integrating shell to take advantage of the database engine associated with Visual Basic™. Further development showed the need to add conditional queries that would filter unwanted and obviously bad data. Additional aggregate queries were added to improve the performance and efficiency of the ARIES application queries conducted during runtime.

### **3. Data Extraction Queries**

As each query was developed, it was first tested in a stand-alone mode and then implemented into the data extraction process. The Administrator Extract Queries Screen is shown in Figure 5. These queries are stored in an Access™ table, named Administrator, and identified by the table name it generates for the ARIES data resource

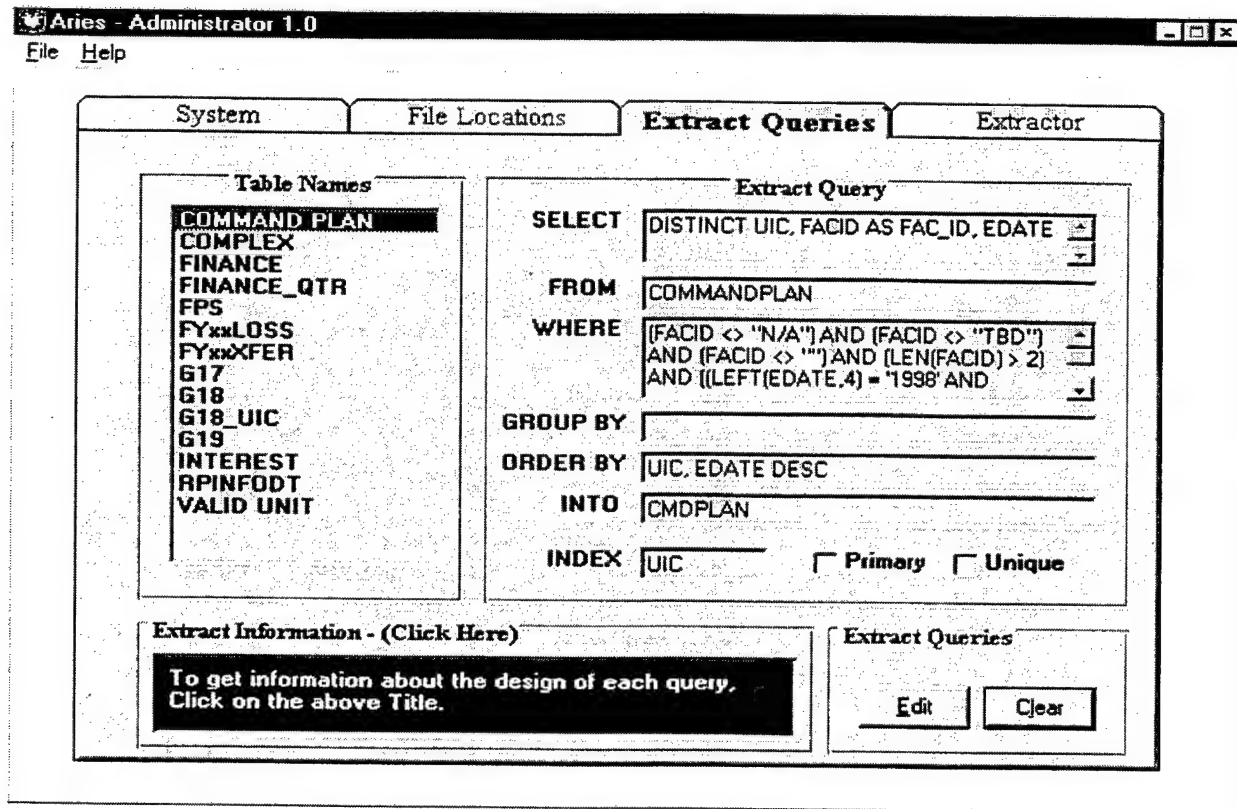


Figure 5. ARIES Administrator Extract Queries Screen

file. These queries can be edited by the administrator to accommodate changes in a source data file or future changes in the application. The extract table structure for each query, as required by the ARIES application, is documented in the Administrator under the Extract Information area and can be reviewed by the administrator. This documents the structure of the table required by the ARIES application so that the administrator can adjust the queries of the extract without affecting the workings of the application.

#### 4. Data Cleaning, Standardizing, and Extracting

The Administrator became a mechanism to transform the original data into a consistent data source for the ARIES application. For this reason, many of the queries that were developed retrieve only the fields and record data required by the business rules for each measure. It also became necessary to standardize the naming of fields that

referenced the same data element because different data files used different naming conventions (e.g., UIC, UIC1, CURR\_UIC). This standardization allowed the application code to remain consistent without concern for the naming conventions used in the source data files and also supported the functionality desired of allowing source files to change without having to change the associated application code. One final task that the Administrator incorporated was a basic cleansing process. Certain values that were identified during initial attempts to query the data as being out of scope or null were removed during the extract. This was accomplished by applying additional criteria to the extract queries.

## **F. CHAPTER SUMMARY**

The overall SDSS architecture of a decision model, a data model, an integrating application system, and a data preprocessor provides simplified access to a set of powerful tools for decision support. These four components generate a working prototype application that is able to complete data analysis in several minutes that would otherwise have taken several groups of individuals many weeks. The decision model is a mapping of the desired decision process into a hierarchy of goals and decision measures that will allow subjective inputs to be achieved for each measure. The data model is generated by developing business rules for each decision measure and identifying source data to answer each of those rules. The integrating application system was designed to bring together the decision model and data model to generate an analysis of a given scenario. The final component, the data preprocessor, is the transformation agent used to prepare and condition the source data for use by the application.

Although it was not a focus of the original project, this system development process gave rise the need for a data warehouse component. As the separate components came together under the original architecture, the system's ability to manipulate data became a limiting factor. The data preprocessor became the transformation agent that was able to remove the required data elements from the operational source files. This preprocessing organized the available information in a format that is optimized to

support the decision process. By cleansing, aggregating, and extracting from the source data files the data preprocessor generated a specialized form of what is termed a data warehouse. Chapter III discusses terms and issues surrounding data warehouses.

### **III. DATA WAREHOUSING AND DATA QUALITY: TOPICAL DISCUSSION**

One of the major consequences of the ARIES project development was the realization that it was necessary to centralize the location of the source data files. This specific user requirement was not identified at the beginning of the project, but rather evolved during the development process as a need to improve system performance during an evaluation session. The process of structuring and creating this centralized data resource resembles some of the current database strategies being used by organizations to take advantage of enterprise wide database information. This chapter provides an introduction to data warehousing, data marts and issues involved in data quality.

#### **A. DATA WAREHOUSING**

The idea of gathering and integrating all the operational information of an organization in one place for the purpose of conducting analysis has been a goal of many information managers. Not until 1990, however, when W.H. Inmon coined the term *data warehousing* was there a formalized architecture or thought process for developing this strategic management tool. Data warehousing, when used properly, will "provide the decision maker of an organization with the timely information necessary to effectively make critical business decisions." [Ref. 4:p. 3] Since 1990 this concept has continued to flourish and grow to the point where today the data warehousing industry is estimated at \$15 billion annually, and 95% of the Fortune 1000 companies have built, or are in the process of building, data warehouses. [Ref. 5:p. 1]

##### **1. Definition**

The term *data warehouse* is a "catch all" phrase that has taken on many different meanings. Michael Brackett defines a data warehouse as "a repository of consistent historical data that can be easily accessed and manipulated for decision support." [Ref. 3:

p. 268-269] Marc Demarest defines a data warehouse as "a consolidation point for enterprise data from diverse production systems." [Ref. 6:p. 1] W.H. Inmon who coined the term initially defines a *data warehouse* as "a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process." [Ref. 7:p. 2] For the purposes of this discussion I will use Inmon's definition and correlate the ARIES project data resource file with this definition.

The concept of "subject-oriented" is based on the change from application-oriented data to decision-support data. Because decision making is the focus, data in a *data warehouse* will be aligned around the major subject areas of an organization whereas operational data will be oriented towards specific business processes it is supporting. Operational application-oriented data are detailed data centered on functional requirements while data for data warehouses will only include data for conducting decision analysis. [Ref. 7:p. 3-4] The ARIES data resource meets these criteria because it contains unit readiness subject data to be used in the decision analysis of site desirability.

Integration as a critical aspect of the data warehouse is an important step that does not always receive appropriate attention. The main focus of the integration process in data warehousing is to obtain consistency throughout the varying legacy databases from which the data are extracted. Consistency can be achieved in many different ways such as standardized naming conventions, measurement of values, encoding structures, and physical characteristics of data. Integration assures that data are stored in a single, globally acceptable manner even if the underlying legacy systems do not do so. [Ref. 7:p. 5-7] Data warehouse integration was a critical objective in creating the ARIES data resource. Naming conventions were standardized (e.g., ZIPCODE, ZIPC, etc. were changed to ZIP), and some data items were manipulated to ensure the consistency of the encoding of the data item (e.g., Nine digit zip codes changed to five digits).

One of the goals of decision analysis is to look at historical data in order to say something about the future. This leads to the need for a time element in the data warehouse to make it an effective tool for decision support. Time variance in a data

warehouse shows up in several ways. First, data warehouses represent data over many different periods of time, encompassing years, year-to-day, months, month-to-day, weeks and days. Second, the index key structure of the data warehouse in all cases maintains an explicit time dimension whereas operational databases are more likely to maintain the time element on an implicit basis. The difference is that the data warehouse will maintain a specific time element as a part of the index key. This is not the case in most operational data files where dates may be associated with the file themselves and not with each data element. Third, the data in a data warehouse are a series of snapshots from the operational database that cannot be updated. [Ref. 7:p. 8-9] In the ARIES context, time is a less important factor than in many data warehouses; specifically, time is visible with respect to the date of the extraction, and therefore the user is aware that the data are assumed to be accurate as of a specific date.

The final defining characteristic of a data warehouse is nonvolatility. This concept rises from the idea that a data warehouse contains a snapshot of the operational data and will not be updated in a traditional sense. The only real functions that happen in a data warehouse are the action of loading the data into the warehouse and any actions accessing that data for the purpose of analysis. This concept provides a stable platform upon which the decision-maker can base decisions. The use of a separate data picture relieves strain on the operational databases from what would otherwise be exhausting analytical queries. [Ref. 7:p. 10-12] In the ARIES project, the Administrator is the agent that allows the data resource for the ARIES project to meet this criterion of a data warehouse. Each extraction of the source files is a snapshot of the source data at that time. One difference between the ARIES Administrator and "standard" data warehouses is that the ARIES data resource file is replaced in whole rather than created as an addition to previous extractions whereas a true data warehouse would build on this historical dataset while extracting and loading new data. The ARIES data resource differs from traditional data warehouses because it is designed specifically to take advantage of the spatial aspects of the underlying data sets. The use of a multi-criteria decision model to



determine the site desirability of an area drove the need to orient data based on its geographic content.

“By any definition, however, a comprehensive data warehouse is much more than archived events equipped with a general purpose front-end query tool.” [Ref. 8:p. 1] As the definition of data warehousing continues to develop, applications of and uses for data warehousing will continue to expand and become more prevalent.

## **2. Applications**

Traditional data warehouses fall into two categories, either Relational Database Management Systems (RDBMS) or Multi-Dimensional Databases (MDDDB). Only in the past few years has data warehousing been viewed as a way for organizations to gain insight about the information embedded in their operational data sets. The necessity for operational data to be reorganized and structured in a data warehouse architecture is driven by the need to maintain acceptable performance and integrity levels in both the operational and evaluational data sets.

An RDBMS is a database system that organizes and accesses data as two-dimensional rows and columns. Data are organized so that related information can be accessed using Structured Query Language (SQL). Data that are linked together with common key values will support a certain level of data integrity, but may create a large amount of overhead at query time depending upon the complexity of the queries required to correlate data elements. Using RDBMSs to support complex analytical processing and decision support has been difficult. The performance of a RDBMS is hindered when it is forced to handle the complex aggregation type queries expected in a data warehouse environment. Each time a query is executed it must aggregate the data that the query is seeking. This sometimes involves millions of records. Until new technologies are developed and tested, RDBMS alone would not be the best choice for a data warehousing project that involves numerous complex queries.

An MDDB is a data base technology that represents multi-dimensional data as aggregations of data in cells that are the intersection of multiple dimensions. A dimension is a table with a single-part key that relates directly to a *fact table* that in turn relates all the dimensions in a star-like structure using a multi-part key. Figure 6 shows an example of a fact table.

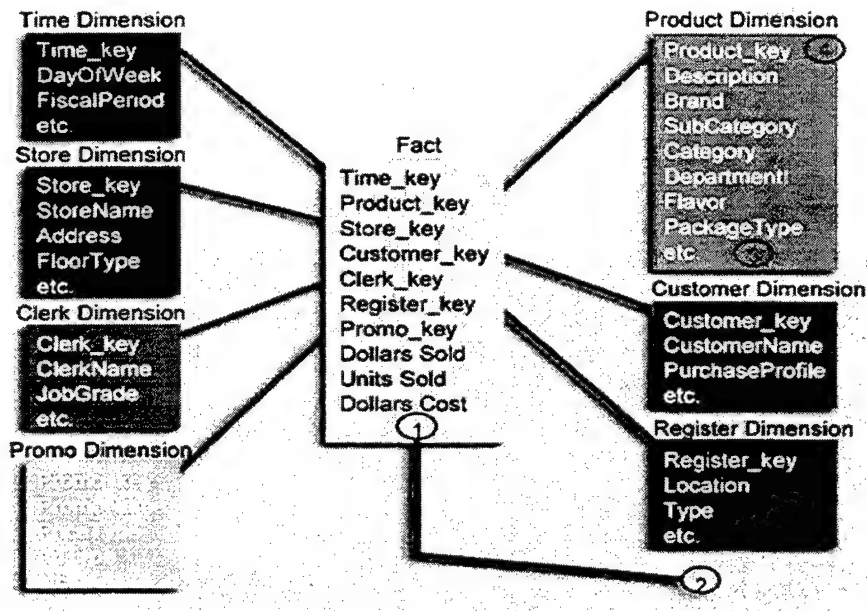


Figure 6. Example Fact Table. [Ref. 9:Figure 2]

The fact table in a MDDB is used to traverse the data across multiple attributes quickly whereas dimension tables contain the actual descriptive data. [Ref. 9:p 2-3] Data in the MDDB will be stored in forms that facilitate the common usage patterns of users. Summarized data that are accessed frequently are preprocessed and made available for the user to query upon demand, unlike the RDBMS that would have to process the query dynamically each time there is a request for that data. This allows for quick retrieval of predefined calculations and efficient results. [Ref. 4:p. 6]

The ARIES data resource file that was generated to support the decision goal of relating TPU readiness to site desirability does not fit either of the two traditional data warehouse types described above. Because the complex queries involved in the ARIES project required aggregating data elements across many data files, the use of a relational

model would have hindered the performance of the evaluation session. The queries for each decision measure, listed in Appendix A, did not require a multi-dimensional analysis and therefore there was no need for an MDDB model. The fact that the 17 different databases represent data from different areas did not allow for the separation of the data into formal dimensions. The geographical nature of the decision goal and all its data files having a spatial aspect qualifies ARIES as a special kind of data warehouse called a Spatial Data Warehouse. The term "*spatially enabled*" is used to describe data that have this spatial or geographical component. Spatial enabling allows data to be related across locations, boundaries and other defined lines that cannot be done easily in the traditional forms of data warehousing.

### **3. Design Concerns**

Marc Demarest defines four fundamental goals of a data warehouse that serve as the basis for complex, forward looking business modeling:

1. To protect production systems from query drain by moving query processing onto a separate system dedicated to that task, and extracting all the relevant information from each production data source at predictable times when off-peak usage patterns prevail;
2. To provide a traditional, highly manageable data center environment for DSS using tools and practices comparable to those used in data center On-Line Transactional Processing (OLTP);
3. To build a Unified Data Architecture (UDA) or Enterprise Data Model (EDM) in the warehouse, so that data from disparate production systems can be related to other data from different production systems in a logical, unified fashion.
4. To separate data management and query processing issues from end-user access issues so that they can be treated as distinct problems. [Ref. 6:p. 4]

These goals provided the foundation for organizations to begin leveraging huge amounts of data they have maintained for years to gain a competitive advantage. They also provide a sound basis to begin the development of a data warehouse project but these

goals fall short in anticipating the dynamic nature of contemporary organizational computing. The last two goals prove to be the biggest drawbacks in this respect.

Developing an in-depth UDA or EDM requires extensive resources and time to complete. Many businesses do not have sufficient physical or financial assets to devote to a project that may not deliver results for a relatively long period of time. Meanwhile, the rapidity of changes in business requirements will inevitably cause the EDM to undergo continuous renovation. As a result, these renovations may very likely be costly without providing timely responses to changes in the analysis needs of the user communities.

The fourth goal of a data warehouse focuses on the data management and querying process, and maintaining these functions separate from the access available to the end-user. This goal is based on the need to perform these large-scale functions in a mainframe based application environment. The performance of current client/server desktop systems has put computing power more directly in the hands of the end-users who can handle portions of these tasks. The ability of the user to manipulate and analyze data directly is required in today's dynamic business environment. [Ref. 6:p. 4]

These shortcomings in data warehouse design architecture gave rise to a more flexible and less expensive solution to organizations' data analysis needs. In 1991 the Forrester Research firm declared data warehousing dead and replaced it with a term they called *data marting*. [Ref. 6:p. 5] The next section will discuss the differences between data marting and data warehousing.

## **B. DATA MARTS**

The terms themselves suggest that the difference between a data warehouse and a data mart would be in the size of data maintained. The difference in size may be true in most cases but more significant differences lie in the application and implementation of the project. Table 3 highlights some of the major differences between a data warehouse and a data mart.

### **Data Warehouse and Data Marts: What's the difference?**

<b>Typically</b>	<b>Data Warehouse</b>	<b>Data Mart</b>
-it addresses:	many subject areas, perhaps the entire enterprise	a subject area
- it is sized at:	GigaByte(GB) to TeraByte(TB)	MegaByte(MB) to Low GigaByte(GB)
- it is accessed by:	business analysts and front-line users	business analysts and front-line users
- it is implemented in:	years	months
- it costs:	\$ millions	\$ tens or hundreds of thousands

**Table 3. Data Warehouse, Data Mart differences. [Ref. 10:p. 9]**

#### **1. Definition**

A Data Mart is a decision support database application that provides decision-making solutions for a narrowly specified group of knowledge workers. The data mart focuses on the needs of the knowledge worker and discounts the underlying production systems in an effort to provide a DSS solution for the workers. This focused approach is achieved by keeping the data mart oriented to one subject area versus the multiple organization wide approach of a data warehouse.

Data marts are more appealing to the business community today because of the reasons mentioned in Table 3 (e.g., size, implementation time, and cost). The smaller size of the data mart compared to the data warehouse allows the information to be available to more users in the distributed desktop environment that characterizes today's business world. [Ref. 11:p. 1-2] This philosophy allows the use of systems that are already in place on decision-makers' desks to conduct detailed decision analysis without the requirement of investing in large amounts of hardware.

The lag time between implementation of a project and output of some useful results is an important issue in the overall success of a project. Josh Bersin, Group Director of Data Warehouse Solutions at Sybase, Inc. indicates that a data mart must

deliver results in the first 90 days. [Ref. 11:p. 2] Because a data mart solution can be designed and implemented in a fraction of the amount of time, i.e., months versus years, it is able to adjust more rapidly to the changing business environment. Data marts should be designed with the concept of expandability in mind because, as users explore the information available, they will want to look at the data in ways for which it was not originally intended. The capacity of the data mart to be flexible and adjust to the user provides the additional feature of scalability.

The ability of an organization to implement a data mart quickly using existing hardware infrastructure provides an immediate cost benefit. In today's business world where every dollar expended is scrutinized closely, it is important to provide business solutions that offer a competitive advantage at a minimum cost. Data marts provide this advantage in their specific subject area. It is important for the organization to ensure that data marts are not built in a vacuum and that each data mart is designed with the enterprise wide data model in mind. This will prevent the proliferation of stovepipe systems. [Ref. 11:p. 1-2]

Marc Demarest first discussed the concept of integration of data marts across the organization in 1993 when he recommended his solution to the enterprise-wide decision support problem. Instead of using only a data warehouse or data mart he recommended the use of a hybrid data architecture. Demarest's thinking was ahead of its time and he was the target of some scathing criticism for suggesting the combination of the two philosophies. In his article "Building The Data Mart", he laid down an architectural model for combining a single warehouse and multiple data marts into one integrated enterprise decision tool that has become one of the most popular designs of enterprise-wide decision support. [Ref. 6:p. 1]

The ARIES project data resource in many ways fits the definition of a data mart. It was implemented in a matter of months, developed on a limited budget, and was designed to support a specific decision process for front-line users. However, in other ways it does not fit the traditional form of a data mart. The ARIES data resource file is a

collection of resources that has been aggregated and manipulated to take advantage of the spatial aspects of each data set. The data files are not maintained in one of the traditional formats of a data mart (i.e., relational or multi-dimensional), but rather are maintained as separate tables that will provide data for each of the twenty decision factors in the most expeditious way. The need to gain performance speed during the querying process forced the use of many geographic queries that quickly localize data by the spatial elements that are already present in the data sets.

## **2. Applications**

Data marts fall into similar categories as the data warehouse that are based on the intended use and types of data that are to be manipulated. The two categories are based on the same design principals as discussed earlier in this chapter for data warehouses, multi-dimensional and relational. A choice between these two design architectures is based on the type of analysis to be done as well as the type of data to be analyzed.

MDDM data marts are used to look analytically at the same data in different ways. They maintain large amounts of numeric data such as sales data. Once the data are loaded, either from the data warehouse or from external sources, it is maintained in a very structured framework. MDDM data marts are most effective for analyzing numeric data in an ad hoc manner. This approach to analytical processing for decision support is called on-line analytical processing (OLAP). [Ref. 12:p. 4]

The relational data mart uses a form of analysis processing known as Relational On-line Analytical Processing (ROLAP). ROLAP data marts support a much wider range of purposes for numeric and textual data and therefore allow for the use of a more general purpose decision tool than the MDDM counterparts. They provide, through the use of relational technology, the ability to conduct both disciplined repetitive queries and ad hoc usage. The data mart concept has its foundation in providing the knowledgeable user with the decision support tool that fits their needs and provides access to just the data that the user needs to see.

### 3. Design Concerns

In designing a data mart the philosophies are founded on the idea of an application being user-oriented in nature. The system is designed to provide the smaller set of users with the exact data set they are going to be using versus an enterprise wide set of data for possible decision concerns. This concept provides for a somewhat different set of processes to be conducted during the design phase. Marc Demarest identified four distinct processes: [Ref. 6:p. 8]

- Extract all data relevant to the business decision-making of the groups of knowledge workers
- Store the resulting data sets in one location: the data warehouse.
- Create a unique cut or series of cuts of the data warehouse for each knowledge worker community. These are the "data marts".
- Supply the decision-support tools appropriate to the knowledge workers' style of computing.

The extraction process involves the translation of the data to standard formats, scrubbing the data for anomalies, and copying only the data elements required for decision-making. This extraction process creates a subset of the operational data set known as a "cut." Storing the data in one location provides a big picture of the business for the major processes in the organization. Because different users throughout an organization apply different aspects of the data it is important to provide each group of users with a specific cut of the data warehouse for their use. This is done with use of data marts. Finally and most importantly, decision-support tools must be provided that match the skill sets of each group of users. If the user is unable to access the resource, it is not an asset but a liability.

The ARIES project conforms to these four distinct processes. All pertinent data are extracted to a single location. The extracted data are intended for use by a specific group of users and in the format that USARC designed. The data are provided through a powerful decision-support tool that provides extensive functionality for the common user as well as detailed analysis capabilities for the knowledgeable user.



Data warehouses and data marts have the same foundation. The data they represent is the detailed data maintained in operational transactional data files. Therefore, the quality of the data retrieved from the data warehouse or data mart is directly related to the quality of the data in the underlying operational data sets. The next section will discuss the issue of data quality as it relates to the data warehouse environment.

### C. DATA QUALITY

The success of a data warehouse or data mart project can not be insured by the best user interface or the newest database technologies if the underlying data is incorrect. The quality of the data obtained for use in any decision support tool becomes a hazard that the users must recognize. It must be managed during the design, development, and implementation phase of any project. Consider the following examples:

- Inaccurate data related to categorization of bank customers resulted in erroneous risk exposure estimates, leading the bank to believe it was more diversified than it was. When the oil market softened in Texas, banks having a large number of Texan accounts suffered a major loss because of the inaccurate representation of risk. [Ref. 13:p. 1]
- A senior level military officer was defending the defense budget before the U.S. Senate. When questioned about a discrepancy in the number of authorized officers shown in the proposed budget versus the congressional numbers, no one could explain the discrepancy. That day the military lost 2,500 authorized officers it needed because Congress liked the lower number. It turns out that a data timeliness problem within one of the data warehouse source systems were the reason for the discrepancy. [Ref. 13:p. 2]

These examples show the necessity of evaluating the value of the data that a decision-maker is using to make important decisions. This section will define *data quality*, identify the attributes that make up the quality of data, and discuss the elements that have the greatest effect on data integrity.

## 1. Definition

With the operative word in data warehousing and data marting being "data," the adequacy of the underlying data used to build the data warehouse must be determined. "*Data Quality*" is the term used to identify and manage the effects of inadequacies of the data in a decision-support environment. Defining data quality and maintaining a level of data quality is a difficult task. It is a common theme throughout the literature that determining the quality level of data used for decision-making is important to the eventual success of any data warehouse project.

Data Quality is defined by Ken Orr, of the Ken Orr Institute, "as the measure of the agreement between the data views presented by an information system and that same data in the real-world." [Ref. 14:p. 2] Richard Wang and Yair Wand define data quality as "a multi-dimensional concept made up of dimensions like accuracy, completeness, consistency, and timeliness." [Ref. 15:p. 87] Michael Brackett states that data quality is an indication of how well data in the data warehouse meet with the business information demand and includes data integrity, data accuracy, and data completeness. [Ref. 3:p. 144] Duane Hufford says "data quality is the state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use." [Ref. 13:p. 1]

It is easy to identify several common elements in the attributes that all of these individuals believe make up data quality. For the purposes of this discussion I will use the definition that is published by the Defense Information Systems Agency (DISA) in the DoD Guidelines on Data Quality Management that identify six characteristics of data quality: accuracy, completeness, consistency, timeliness, uniqueness, and validity. [Ref. 16:p. 2] Table 4 gives a description and an example of each of these six characteristics. The application of the use of these six characteristics provide a sound foundation for any organization to begin the process of identifying a confidence level for the quality of data that are being used for decision-making. These characteristics are present in most legacy databases in some form, which makes any data warehouse to which this information is migrated susceptible to the same types of errors present in the legacy system. "One data

<b>Data Quality Characteristics</b>	<b>Description</b>	<b>Example Metric</b>
Accuracy	A quality of that which is free of error. A qualitative assessment of freedom from error, with a high assessment corresponding to a small error.	Percent of values that are correct when compared to a the actual value. For example, M=Male when the subject is Male.
Completeness	The degree to which values are present in the attributes that require them.	Percent of data fields having values entered into them.
Consistency	A measure of the degree to which a set of data satisfies a set of constraints.	Percent of matching values across tables/files/records.
Timeliness	It represents the degree to which specified data values are up to date with the real-world.	Percent of data available within a specified threshold time frame(e.g., days, hours, minutes)
Uniqueness	The state of being the only one of its kind. Being without an equal or equivalent.	Percent of records having a unique primary key.
Validity	The quality of data that is founded on an adequate system of classification and is rigorous enough to compel acceptance.	Percent of data having values that fall within their respective domain of allowable values.

**Table 4. Data Quality Characteristics. [Ref. 16:p. 2]**

manager for a large company reported that fully 60% of the data that was transferred to their data warehouse failed to pass the business rules that the systems operators said were in force.” [Ref. 14:p. 6]

Because the recipients of this “dirty data” are the resources that organizations are using to make “fact based” decisions from, it is important to understand data quality as it relates to the project. Organizations create data warehouses and DSSs to avoid making inaccurate assumptions about their business. They should then not make assumptions about what makes up the data set used to load the data warehouse. [Ref. 17:p. 1] Creating a plan to improve data quality is part of understanding that legacy data itself will not meet the quality standards required to make decisions.

## 2. Improving Data Quality

The process of improving data quality is an incremental process that can be conducted in all phases of a project from birth to implementation and beyond.

Although database consultant Ken Orr has called data warehouses the sewage treatment plant of enterprise data, this is not the objective of data warehousing. It is, unfortunately, the unintended result of loading legacy data that has not been subjected to data-quality improvement. [Ref. 18:p. 1]

To improve the quality of data in any system you must first determine the baseline quality of the current data set must first be determined. This is done by comparing actual data instances against the established rule sets that have been established. The rule sets will become the metrics like those shown in Table 4 and will be documented in the meta-data of the data files for future reference. If these rules are not documented, the first step in improving data quality is to document the business rules that the current data represent. An important part of this initial assessment is to identify the responsible stakeholder or stakeholders and to get them involved in the improvement process. Once the data quality baseline assessment is complete the next step is to determine and document the level of quality required by the intended business use.

There will be different levels of quality required for data depending on the intended use. Ken Orr in his discussion of Data Quality and Systems Theory states that:

No serious information system has a data quality of 100%. The real concern with data quality is to insure not that the data quality is perfect, but that the quality of the data in our information system is *accurate enough, timely enough, and consistent enough* for the organization to survive and make reasonable decisions. [Ref. 14:p. 3]

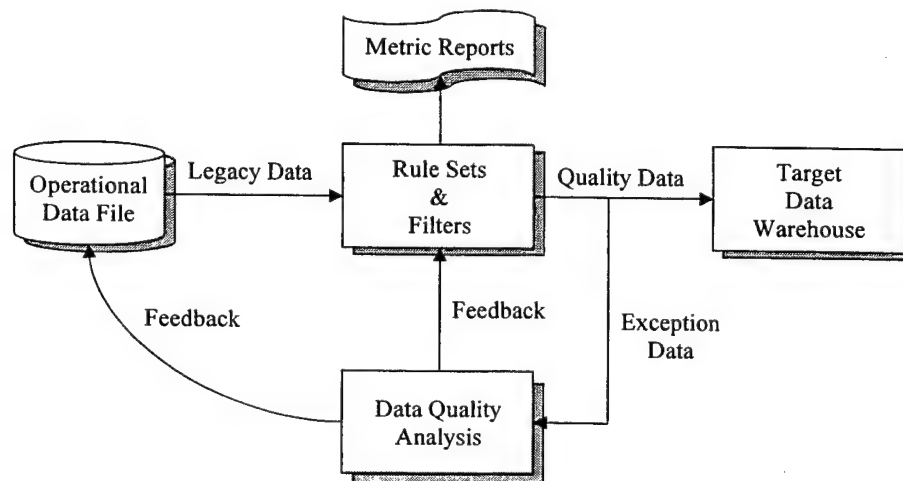
A system that is used to make life threatening decisions will have need for higher quality data than system used to identify the placement of a new facility. This required quality level or the "*enough*" that Orr speaks of can only be determined by the users of the system. Once the user determines the level of quality required, it becomes the goal of the project to attain that level. Part of attaining that level is to have in place the mechanisms to measure and maintain that quality over the life of the project. Where possible this

should be done at the origination point of the data instance. Larry English states that “to improve warehouse data quality you must improve the business processes that produce the data.” [Ref. 16:p. 4] If this is not possible, quality controls must be put in place in the migration path of the data from the operational source to the data warehouse or data mart. This process is known as “*data cleansing*.” Because the source files that were used in the ARIES project are not in direct control of the customer, the process of *data cleansing* was used extensively in the development of the ARIES data resource file.

### 3. Data Migration

The operation of moving data or loading it into the data warehouse is not a simple task and requires substantial planning. As pointed out previously in this chapter, if you simply move the data from the operational system into the warehouse you are moving the data problems as well. The process of migrating the data therefore becomes a point at which problems with the legacy data can be identified and possible solutions can be developed.

The migration process involves a method to transfer the data to the target data warehouse, transformation of the source data into the data warehouse architecture, and a method to clean or scrub data problems. Figure 7 is a diagram of the migration process that shows the flow of legacy data into the target data warehouse.



**Figure 7. Data Migration Process**

The perfect migration process would involve continuous measurement of the data being transferred against the business rule sets that is in place and has the ability to provide feedback in the form of a control system. The concept is discussed both by Ken Orr in his paper on Data Quality and Systems Theory [Ref. 14] and in the DoD Guidelines for Data Quality that is published by DISA [Ref. 16].

The function of this migration process is handled by the ARIES Administrator in the ARIES project. The Administrator maintains the business rules that USARC developed and conducts the transfer of data from the source files to the target data resource file. It does not contain the ability to ascertain the quality of the data being transferred or measure that quality against any form of metrics. This process must be conducted outside of the automated migration process provided in the ARIES SDSS project.

#### **D. CHAPTER SUMMARY**

The philosophies and technology involved in the data warehousing process are currently in their infancy. The concept is growing rapidly as is shown by its wide acceptance in the leading business communities. As long as the concept of subject-oriented, integrated, time-variant, and nonvolatile data collections continues to provide businesses with a competitive advantage, the data warehouse will be at the center of the enterprise decision-making tool set. Data marts have already proven to be an acceptable way to provide a specific cut of data to a particular group of users with the intent of providing greater accessibility. The use of data warehouses and data marts in a common architecture will provide even greater access to the enterprise data. Greater access will give rise to new uses for that data and allow the users to maximize their use of the available data.

The ability of an organization to take advantage of "spatially enabled" data is a new concept that requires much more research and development. Data that are spatially enabled will begin to link data that otherwise had no common link and will allow the

organization to fine tune their decision-making on a new level. The ARIES project is a working prototype of an application that has made every attempt to maximize the spatial elements of the source data. This has allowed the automation of a decision process previously thought to be too complex to be handled in a client/server environment.

One aspect of the data warehousing and data marting concept that is consistent, as long as the data being warehoused is legacy data, is the need to determine the quality level of the source data sets. "Data quality" is defined by six characteristics of data: accuracy, completeness, consistency, timeliness, uniqueness, and validity. The quality of the data that are being used can be improved during all phases of a project. The improvement process involves identifying ownership of the source data, conducting a data quality baseline assessment, and determining the required level of quality for the system. For the quality improvement plan to be effective it must provide a mechanism of providing feedback to the data systems for the life cycle of the system. This continued quality assessment commonly takes place during the migration process. This process allows the data to be transferred, transformed, and scrubbed to meet the requirements of the data warehouse architecture.

Chapter IV will discuss the how the quality level of the source data files for the ARIES SDSS project were determined and identify the anomalies that were identified.

## IV. ARIES SDSS APPLICATION: PROBLEM EVALUATION

The ARIES SDSS prototype was developed in direct response to the requirements set forth by the intended users at USARC. The final prototype is a result of continuous adaptations to changing user requirements, usability improvements, and solutions to system design and implementation problems. Problems were encountered in several areas during the development process; business rule development, query performance, and data anomalies. This chapter discusses specific examples of problems from each of these areas as well as the methods used to overcome them.

### A. BUSINESS RULE DEVELOPMENT PROBLEMS

The business rules of the SDSS are the links that connect the hierarchical decision goals with actual data instances in source data files. The business rules of the ARIES project decision measures are described in detail and listed in Appendix A. An example of an ARIES SDSS business rule is the following:

<u>Area Loss Rate</u>	Is equal to the number of losses to units in the area during the previous fiscal year divided by the total number of reservists currently assigned to those units. Losses are determined by counting the entries in the FYxxLOSS file where the data element "TRMN" equals "LOSS" that are associated with each UIC in the area. The total number of assigned reservists is determined by counting all of the personnel records in the G18CWE file associated with each UIC in the area. "In the area" is defined as within a 50 mile radius of the proposed site.
-----------------------	--

Some of business rules documented during the project development phase were flawed in their initial assumptions. The problems encountered fall into two main categories: errors in logic and rules not support by data. The next sections will discuss examples of these problems.



## 1. Errors in Logic

An example of a logic problem occurred with the requirement by seven of the 20 measures to identify all units “in the area” of a proposed facility. The business rule developed by the expert panel at USARC identified a file called COMMAND PLAN as the best source from which to obtain this information. This was based on the assumption that COMMAND PLAN contained entries for every unit and facility in the Army Reserve. After reviewing the actual data entries in the file, however, it was found that the file actually contained multiple entries for each site, including information on closed and proposed sites as well as active sites. It was therefore impossible to identify a list of valid units or facilities solely from this data source. When the system was run under this assumption the program failed because of multiple entries with same values in a key index field.

It was then determined that this information must be derived from a complex query that matched the entries in COMMAND PLAN against the G17 source file that lists basic facility information (Table 5). This query, referred to as VALID\_UIC, creates a table of UICs that is created and only stored in the ARIES SDSS data resource file.

Because the COMMAND PLAN source file contains historic, present, and future entries

```
VALID_UIC
SELECT  UIC, FAC_ID, UnitName, City, State, Zip
FROM    G17Natl
WHERE   G17Natl.UIC = ANY (SELECT CMDPLAN.UIC
                           FROM CMDPLAN)
```

**Table 5. VALID UIC Query**

for each site, a filter query was added to the extraction process in the Administrator to load only data instances from COMMAND PLAN that are valid over the next 13 months, as shown in Table 6. The combination of these two queries solved the problem of identifying valid Army Reserve units.

```

CMDPLAN
SELECT    DISTINCT UIC, FACID AS FAC_ID, EDATE
FROM      COMMANDPLAN
WHERE     (FACID <> "N/A") AND (FACID <> "TBD") AND
          (LEN(FACID) > 2) AND ((LEFT(EDATE,4) = 'CCYY'
          AND MID(EDATE,5,2) <= "MM") OR (LEFT(EDATE,4)
          <= 'CCYY'
ORDER BY  UIC, EDATE DESC
INTO      CMDPLAN
INDEX     On UIC As UIC

```

Note: Application automatically adjusts the dates to obtain a 13-month window.

**Table 6. COMMAND PLAN Filter Query**

## **2. Rule not Supported by Data**

An example where underlying data did not support the business rule was uncovered while determining the value for backlogged maintenance actions of a facility, Measure #1 in Appendix A. The original business rule applied a criterion of totaling only the "K-account" unfunded requests, identified by the fund code of "BMAR". After reviewing the data file and attempting to implement this business rule, it was discovered that only a small number of facilities had any entries with the BMAR fund code. Because a value of zero for backlogged maintenance receives the maximum utility, as shown in Appendix A, this error would have seriously overestimated the contribution of backlogged maintenance to site desirability and potentially biased the outcome of the decision model significantly. This problem was solved, after discussions with USARC, by totaling all the unfunded requests without concern for the fund code.

## **B. QUERY PERFORMANCE PROBLEMS**

The user did not care how long the evaluation took as long as it was an automated process. The ARIES SDSS project requirements offered little insight about the expected computer execution time a site evaluation would take. During the beginning phases of prototype development, query times in excess of one and half hours were common. This long evaluation time, though still many times faster than the current manual process, was

considered unacceptable for an automated decision support implementation by the development team. Furthermore, the source files in use during the beginning phase only contained data for the state of Pennsylvania. Extending the data sets to the national level promised some truly staggering execution times. Efforts were undertaken immediately to streamline the lengthy query by focusing on two areas of the querying process: (1) the use of geo-queries in place of standard SQL queries and, (2) aggregation of detailed information into smaller data sets.

### **1. SQL vs. Geo-Query**

After reviewing each measure in detail, it was determined that 14 of the 20 decision measures were dependent on a spatial element query (e.g., Number of Reservists within 50 miles). Because the application was using a GIS system already, a natural course of action was to leverage the powerful geocoding abilities of MapInfo™ to conduct spatial queries as a way to reduce the overall query time. The MapInfo™ queries executed three to four times faster than conducting the same query through standard SQL. Geocoding a source file allows MapInfo™ to localize the desired records and only look at a subset of the data file. A counterpart SQL command, on the other hand, would attempt to match each record in the data file. The processes of passing the queries to MapInfo™ reduced the evaluation time from hours to less than ten minutes.

Another obvious advantage to using MapInfo™ to conduct other queries involving a list of items with a specified area of a geographical location. This feature was also used to determine the following lists for use elsewhere in the application: units within 50 miles, facilities within 50 miles, Army reservists within 50 miles, members of the National Guard within 50 miles, Individual Ready Reserve (IRR) individuals within 50 miles, and a list of zip codes within 50 miles.

## 2. Detail vs. Aggregation

One of the factors that led the development team to the incorporation of data warehouse concepts in this project was the requirement which surfaced during development to improve query performance. As discussed in Chapter III, a data warehouse is an optimized data store used to provide data in a structure or format that will maximize the performance of the DSS tool set. Aggregating the detailed information in source files into summarized tables is one of the techniques use to increase performance. The use of aggregation to improve performance was an idea that was used heavily in developing the ARIES data resource. Because a number of the queries required counting the records that match a particular criterion, the concept of aggregation provided an obvious advantage.

An example of the benefits of aggregation can be seen in the counting of the number of individual reservists assigned to a list of units. The original process would have conducted a complex query that counted the number of entries in the G18CWE personnel file that had a UIC matching any one of the UICs determined to be in the area of the proposed site. The G18CWE data file has in excess of 200,000 records. Matching each entry against a list of any substantial number of UICs took in excess of an hour. The solution was to add an aggregation query, shown in Table 7, to the data preprocessing phase that counted the entries for each distinct UIC and maintained only that total. After implementation of this query the application would then only query the aggregate table for each UIC on the area list and conduct a simple summation query. This form of aggregation was implemented in the extraction process for four of the 17 databases: FINANCE, FYxxLOSS, G18CWE, and RPINFODT. Those queries are listed in detail in Appendix B. This process moved the query time required for the counting process into

```
SELECT UIC, COUNT(UIC) AS UIC_TOTAL
FROM   G18NatI
ORDER BY UIC
GROUP BY UIC
INTO   G18NatI_UIC
```

**Table 7. Example Aggregation Query - G18NatI UIC**

the data preprocessing session and out of each individual evaluation session, reducing site evaluation time by about half.

### **C. DATA ANOMALIES**

The final and most intricate problem area was the quality of the source data. As discussed in Chapter III, the solution a DSS provides is only as good as the data on which it is based. The quality of the source data files for the ARIES SDSS project provided a substantial challenge to the development team.

Early in the development process, the frequency of data values that were missing or null caused multiple error conditions in the applications. A need arose to identify missing data values with a default error value so the application did not have to contend with null values. A value of "-999" was returned for a decision measure as the default error value. Flagging this one error value identified many inconsistencies in the source data files. Additionally, the magnitude of the number of default error values that the system returned became a concern during initial testing by the user. The initial intent of inserting error values was to allow the user an opportunity to enter a subjective value in place of the missing value. This subjective value would hypothetically allow the decision modeler to provide a better approximation of site desirability. Some runs of the application returned so many default error values, however, that the user would have been entering more values for decision measures than the automated application returned, thus defeating the primary purpose of the system. This was deemed an unacceptable condition.

#### **1. Proxy Value Calculations**

It was decided that an interim solution to the number of error values returned was to determine a proxy or default value for each measure. These proxy values could be substituted automatically in place of the "-999" values to allow the decision model evaluation to be conducted without sacrificing authenticity completely. Determining a

value that would provide an accurate representation for a given measure required the calculation of basic descriptive statistics for each measure (e.g., Mean, Standard Deviation, Minimum and Maximum values).

Descriptive statistics could be calculated for 17 of the 20 decision measures for each facility because the business rule did not depend on knowing the Moving Unit. These measures were facility oriented or oriented on the area around the proposed facility independent of any characteristics relating directly to the moving unit. The three measures that could not be calculated in this way because they do depend on moving unit characteristics are: Number of Reassignments from the Moving Unit, Available individuals with MOSs of interest from Closing Units, and Available IRR individuals with MOSs of interest.

In an attempt to determine a source for the error conditions, as well as calculate descriptive statistics, a complete evaluation of all possible sites was conducted. This "global" evaluation process allowed the application to be tested to the full extent of the data set and assisted the development team to identify potential problems quickly. The procedure was conducted twice and required the application to run without interruption in excess of a week (using a Pentium 90MHz personal computer). The resulting descriptive statistics for 17 of the 20 decision measures are listed in Table 8. Appendix C contains a detailed listing of the descriptive statistics and frequency data for each decision measure.

## **2. Data Validation**

Because of the enormous amount of missing or null values that the system was returning during initial evaluation sessions, it became necessary to verify and validate the data set in use for the ARIES SDSS prototype. This was required to localize the problem and determine if the problem was with the data or in the application implementation.

Before the data set could be validated, an appropriate range of data values had to be determined for each measure. The limits of the ranges were determined using a rule of reasonableness to identify values that would adversely affect the evaluation process.

Measure	Obs (N)	Min Value	Max Value	Mean	Std Dev
1. Facility Backlogged Maint.	1,205	0	11,979,371	448,131	837,391
2. Facility Operating Cost	1,251	0.0	293.5	3.1	9.7
3. Facility Age	765	0	1,677	295.1	173.3
4. Facility Condition	1,251	N/A	N/A	N/A	N/A
5. Facility Owned	1,319	N/A	N/A	N/A	N/A
6. Competition	1,300	18	20,759	4,116.3	3,960.0
7. Area Drill Attendance	1,300	0.20	0.82	0.58	0.06
8. Area Loss Rate	1,325	0.00	0.86	0.32	0.11
9. Area Transfer Rate	1,300	0.00	1.84	0.27	0.20
10. Area Average Manning	1,325	0.00	1.94	0.86	0.20
11. Distance to Recruiter	1,325	0	7,619.9	18.2	287.7
12. Area Available Closing Unit	819	1	504	75.2	114.1
13. IRR Available	1,315	1	3,497	395.8	658.9
14. Area Recruit Market	1,316	253	214,738	33,189.9	41,290.4
16. Distance AMSA	1,325	0	7,619.9	42.4	289.1
17. Distance ECS	1,325	0	5,290.9	268.1	510.1
18. Facility Weekends Used	1,320	0	3	1.6	1.0

Note: Measures 15, 19, and 20 are dependent on the Moving Unit

Total Number of Facilities (N): 1325

**Table 8. Descriptive Statistics for Decision Measures**

Consideration was given to the following areas; (1) the range of values returned during the evaluation process, (2) expected values based on the Yield Curves, and (3) common sense (i.e., a value of zero for Facility Age was not considered reasonable). The valid ranges for each measure are listed on Table 9.

As indicated in Table 9, major problems with at least six of the 20 decision measures were identified. The validation was conducted on the list of 1523 available facilities. The review identified a serious problem with files which contained missing values; in some cases the files were missing as much as 57% of the values (Area Drill Attendance). A detailed review of the data files in question determined that the files did not contain information for any state other than Pennsylvania. This problem was a result of the user requirements from a prototype for Pennsylvania data to a national data set. This change occurred during the development process without the data files being

Measure	Missing Values (%)	Out of Range (%)	Potentially Valid (%)	Valid Range
1. Facility Backlogged Maint.	14.8	2.5	82.7	$0 < x_1 < 20M$
2. Facility Operating Cost	8.5	24.4	67.1	$0 < x_2 < 100$
3. Facility Age	49.6	0.1	50.2	$x_3 > 0$
4. Facility Condition	8.5	0.0	91.5	$x_4 = G \text{ or } A \text{ or } R$
5. Facility Owned	1.0	0.0	99.0	$x_5 = Y \text{ or } N$
6. Competition	2.7	0.0	97.3	$0 < x_6 < 21,000$
7. Area Drill Attendance	57.1	8.4	34.5	$0 \leq x_7 < 1.0$
8. Area Loss Rate	2.4	60.3	37.2	$0 < x_8 < 1.0$
9. Area Transfer Rate	2.7	56.5	40.8	$0 < x_9 < 1.0$
10. Area Average Manning	0.0	3.2	96.8	$0 < x_{10} < 1.5$
11. Distance to Recruiter	0.0	0.7	99.3	$x_{11} > 500$
12. Area Available Closing Unit	38.1	0.0	61.9	$x_{12} \geq 0$
13. IRR Available	1.3	0.0	98.7	$x_{13} > 0$
14. Area Recruit Market	1.2	0.0	98.8	$x_{14} > 0$
16. Distance AMSA	0.0	0.7	99.3	$x_{16} > 500$
17. Distance ECS	0.0	11.6	88.4	$x_{17} > 500$
18. Facility Weekends Used	0.9	0.0	99.5	$x_{18} > 4$

Note: Measures 15, 19, and 20 are dependent on the Moving Unit.

Total Number of Facilities (N): 1523

**Table 9. ARIES Measures Analysis Statistics - Run #1**

updated to match the new national search requirements. Databases containing nationwide information were used to conduct a second analysis and evaluation session as discussed later in Table 10.

A problem was also identified when using a ratio as the measuring metric. A ratio does not reflect the magnitude of the underlying values used to obtain that ratio. In the case of the value for Area Drill Attendance, for example, the application returned a value of 0.8 for a facility, which would be considered within the expected range. The value for Area Drill Attendance involves calculating the ratio of reservists meeting satisfactory drilling requirements to the total number of reservists required to drill at a facility. The application was constructed to store the interim values of DRILL\_SAT and DRILL\_TOTAL, which were calculated to be 4 and five 5 respectively, resulting in a



ratio of 0.8. However, these numbers did not match the total number of reservists actually assigned which was 2496. This wide disparity was traced back to the same problem leading to 57% of the facilities having missing values for Area Drill Attendance. This problem was resolved when the FINANCE file was updated to reflect entries on a national basis. However, it is still possible for a measure represented as a ratio to hide potential data problems. A good strategy in this case is to display the basic values that comprise the ratio in addition to the ratio itself. Five of the decision measures in the ARIES SDSS project use ratios: Area Drill Attendance, Area Loss Rate, Area Transfer Rate, Average Area Manning, and Reassignments. A detailed description for the calculation process for each measure is shown in Appendix A.

One potentially deceptive measure for which values calculated during the analysis may provide false feedback is Area Available Closing Units. This measure is used to determine the number of available reservists from units that are scheduled to close in the area of a proposed site. Because the number of closing units is small in comparison to the number of active units, not all facilities will have units scheduled for closing within 50 miles. However, the application will return a default error value for this case that should not be considered an error. In this case a value of zero could be the valid answer. Because it is difficult to distinguish between missing data or the fact that there may be no closing units, the application currently does not compensate for this situation and the correction is left to the user.

A second analysis session was conducted on a list of facilities known to be active Army Reserve facilities in the continental United States. The original list of 1523 facilities was pared down by a total of 198 sites to a total of 1325 by removing facilities in remote locations, facilities marked as not existing, and facilities marked as temporary. Appendix C contains the frequency data and descriptive statistics for each measure that was produced by this second analysis run. Table 10 shows the validity analysis statistics for each measure.

Measure	Missing Values (%)	Out of Range (%)	Potentially Valid (%)	Valid Range
1. Facility Backlogged Maint.	9.1	1.7	89.2	$0 < x_1 < 20M$
2. Facility Operating Cost	5.6	18.3	76.1	$0 < x_2 < 100$
3. Facility Age	42.2	0.2	57.6	$x_3 > 0$
4. Facility Condition	5.6	0.0	94.4	$x_4 = G \text{ or } A \text{ or } R$
5. Facility Owned	0.5	0.0	99.5	$x_5 = Y \text{ or } N$
6. Competition	1.9	0.0	98.1	$0 < x_6 < 21,000$
7. Area Drill Attendance	1.9	0.0	98.1	$0 \leq x_7 < 1.0$
8. Area Loss Rate	0.0	2.1	97.9	$0 < x_8 < 1.0$
9. Area Transfer Rate	1.9	1.1	97.0	$0 < x_9 < 1.0$
10. Area Average Manning	0.0	2.3	97.7	$0 < x_{10} < 1.5$
11. Distance to Recruiter	0.0	0.1	99.9	$x_{11} > 500$
12. Area Available Closing Unit	38.2	0.0	61.8	$x_{12} \geq 0$
13. IRR Available	0.7	0.0	99.3	$x_{13} > 0$
14. Area Recruit Market	0.6	0.0	99.4	$x_{14} > 0$
16. Distance AMSA	0.0	0.1	99.9	$x_{16} > 500$
17. Distance ECS	0.0	10.6	89.4	$x_{17} > 500$
18. Facility Weekends Used	0.4	0.0	99.6	$x_{18} > 4$

Note: Measures 15, 19, and 20 are dependent on the Moving Unit.

Total Number of Facilities (N): 1325

**Table 10. ARIES Measures Analysis Statistics - Run #2**

Five of the measures in this run fall below a 90% potentially valid level. Three of these measures (Facility Backlogged Maintenance, Facility Operating Cost, and Facility Age) are facility related and are the result of missing data or, in the case of facility operating cost, values of zero.

The missing values for Area Available Closing Unit, as discussed above, are a result of proposed sites not having any closing units within the geographic area. In this case it would be acceptable for a site to return a null value for this measure. The values for the Distance to ECS measure are a result of the Army Reserve having only 30 ECS sites nationwide. As a result some sites will have a distance of greater than 500 miles to the nearest ECS site. This "out of range" distance value will be interpreted by the decision model and assigned the minimum utility value based on the shape of the yield

curve. Hence, these invalid range problems for these two variables will not adversely affect the decision model.

Missing or null values in the data files are identified in the application by a default error value which can be changed easily by the user in order not to affect adversely the outcome of the evaluation. The values that are out of range, on the other hand, can have a direct, negative impact on an evaluation if they are not identified and corrected. Consider Facility Operating Cost measure as an example. Of the 18.3% of the values that were found to be out of range, all but two were equal to zero. This is a concern in the evaluation process because a value of zero will receive the maximum utility during the decision analysis phase. It is true that a closed facility or a proposed facility would not have a current value for the operating costs but in the case of an active facility this value should be something greater than zero. Currently this situation requires the knowledgeable user to intervene and apply a reasonable value. This is a particularly insidious example of how incomplete or inaccurate values in source files can filter through the data warehouse into the DSS.

### **3. Data Quality Analysis**

The data analysis conducted for the SDSS application was only concerned with determining data validity. A complete analysis would consider all the characteristics of data quality discussed in Chapter III including accuracy, completeness, consistency, timeliness, and uniqueness as well as validity. Below we present some examples of data problems encountered during the development phase, which relate to these other data quality characteristics.

#### ***Accuracy***

Problem: (Facility Condition - Measure # 4)  
Every value for facility condition is "GREEN". It is unreasonable to expect that no facilities would be coded either AMBER or RED, therefore it seems quite unlikely that these values match the actual condition of each facility.

Solution: These values must be updated by the owner/stakeholder of the operational data file.

### ***Completeness***

Problem: (Facility Age - Measure # 3)  
The source data file is missing 42% of the facilities deemed to be valid from the GEOREF file.

Solution: These values must be updated by the owner/stakeholder of the operational data file.

Problem: GEOREF File  
Zip codes are missing on 96 of the facilities listed in GEOREF, 20 of which are marked as valid facilities.

Solution: These values must be updated by the owner/stakeholder of the operational data file. The zip code is used to geocode these facilities for possible selection as a proposed site.

### ***Consistency***

Problem: Zip Codes are of varying length in all the data files. Some files contain the nine digit zip codes while others only maintain five digit codes.

Solution: In order to query on zip codes and obtain an exact match, the use of five digit zip codes was adopted. This was done during the extraction process by only loading the first five digits of a zip code from all the files.

Problem: UICs are not represented uniformly in all the source data files. Some files use a UIC designation that does not include the letter designating an active unit. Other data files use a parent UIC instead of the UIC of an actual unit. (e.g., unit structure is against the parent UIC, .....AA, whereas the person assigned to a billet is assigned at the platoon level UIC, .....A1)

Solution: Ensured all UICs were six digits in length.

Problem: Data entries for facilities and units in source files do not match the list of valid units.

Solution: Not corrected. Action required by the owners of the source data files.

### ***Timeliness***

Problem: (Area Loss Rate, Area Transfer Rate, Measures # 8 & 9)  
FYxxLOSS File used to determine loss and transfer rate can be as much as twelve months out of date.

Solution: No current solution.

### ***Uniqueness***

Problem: The data files do not have unique indexes. The lack of a unique list of facilities and units has allowed entries in source data files for sites that do not exist.

Solution: No current solution.

These examples are not exhaustive and are only intended to be representative of the problems that exist in the source data files for the ARIES SDSS project. Further analysis is required in the area of the other five data quality characteristics to determine the overall level of data quality that is present in the final application. Because the SDSS is based on data files that are updated on a frequent basis, there is a need to monitor the quality level of the data following each future extraction.

## **D. CHAPTER SUMMARY**

As with any prototype DSS development project, substantial problem areas arose with respect to data quality issues. Three major categories of data quality problems were identified: business rule development, unacceptable query performance, and data anomalies. Though these problems provided a considerable challenge to the development team, acceptable solutions have been implemented in most cases.

The documentation of the business rules provides the building blocks of the SDSS application. In the ARIES application, problems with the business rules manifested as logic errors or lack of data support. Solutions to these problems can be developed for both types of business rule errors. Unfortunately, these errors usually show up during the initial phases of the application development and must be resolved to allow the application to continue with development.

Because the SDSS application is query intensive, the performance level of the application will depend on the ability to carry out each query in the most efficient manner. The ARIES application required performance enhancements in two areas: geo-queries for spatially related queries and preprocessed data aggregation. These relatively simple solutions provided a performance enhancement that reduced the evaluation time for each site by an order of twenty fold, from hours to a matter of minutes.

The validity and quality of the source data directly affects the quality of the resulting evaluation in the ARIES SDSS application. An understanding of the data problems involved in the ARIES project did not become a concern until the application was in full development. Because of the magnitude of the problems, steps were initially taken to localize the source of problems. Through a detailed validation analysis, a large portion of the data anomalies was corrected. Data problems were handled either at the preprocessing stage with filtering queries or else corrected at the source file. The correction of these anomalies on a post facto basis consumed a major amount of time that detracted from development. This opportunity cost of undertaking remediating action could be reduced significantly with prior planning.

The problems that the ARIES SDSS prototype application encountered were in most cases not anticipated. The need to reduce the impact of these problem areas on the application development process requires changes in the initial steps of the SDSS development process. Chapter V discusses the lessons learned from this initial prototype development and proposes several requirements that should be added to the development process in order to identify solutions for these problem areas earlier in the project life cycle.



## **V. LESSONS LEARNED: SDSS DESIGN AND DEVELOPMENT**

The ARIES SDSS project prototype is in final implementation, but it did not get to that point without encountering a number of major problems that could have been avoided with prior knowledge. The lessons learned from this project will be valuable to the development of second-generation SDSS projects.

The majority of the problems with the ARIES project centered on the wide disparity between the initial system concept and the final product. The original goal of the project was to develop a decision model for the TPU readiness issue. However, the final product is a fully functional automated decision support tool. As the ARIES project grew with each new functionality the development process itself received less and less attention. Writing and testing code became the focus. While this is a common scenario for prototype applications, serious problems can arise when a decision is made to take the prototype to full implementation with little planning, which is what occurred in this situation.

This chapter discusses recommended changes to the SDSS development process to avoid the pitfalls that hampered the development of the original ARIES prototype application. These changes include the addition of a data migration plan, refinements to the decision model, data model, and system design, and future considerations.

### **A. SDSS DEVELOPMENT PROCESS**

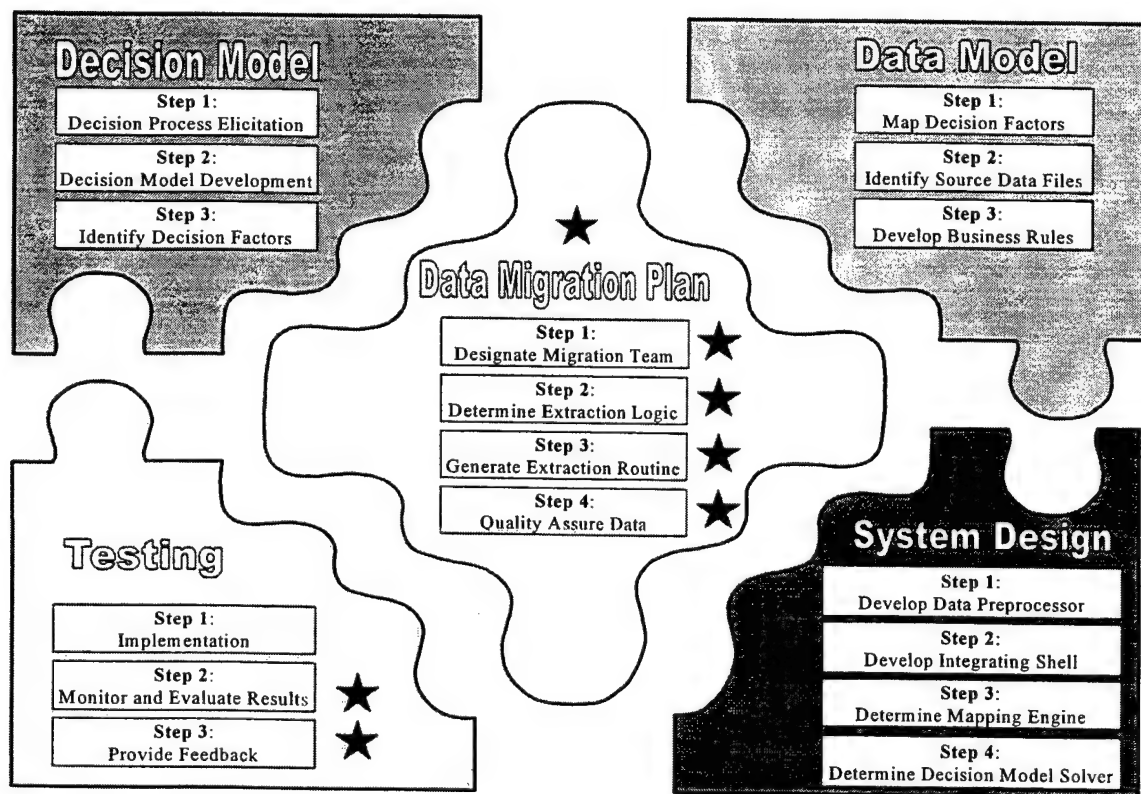
The ARIES SDSS began as a project to elicit an expert system decision model for placement of Army Reserve TPU units using a GIS to display locational decision factors. As the users generated additional requirements, the project evolved into an integrated application data resident model involving the use of a GIS application and a DSS application. Because the original scope of the project centered on the accuracy of the decision model, the retrieval of data and data quality issues supporting that model were not in the original considerations.



This complacency about the underlying data mirrors a recurring theme in the data warehouse development literature. Joe Celko and Jackie McDonald indicate the likely consequences of such oversights.

“Ignore or trivialize problems with the existing data at the start of the project, and that oversight will brutally assert itself when data problems begin to surface as you populate the warehouse from outside data sources, current applications, and legacy data.”[Ref. 19:p. 1]

Unfortunately, this statement was particularly appropriate in the case of the ARIES prototype application. Overlooking the migration of the source data to the application was a major oversight. This error has led us to recommend a revision of the SDSS development process that is documented in Chapter II (Figure 8). The major difference from the original process is the addition of a data migration plan (DMP) discussed below.



★ Identifies recommended changes to the original development process.

**Figure 8. Recommended SDSS Development Process**

## **B. DATA MIGRATION PLAN**

As the ARIES project developed, the data model evolved from merely querying source files into a specialized form of a data warehouse. The importance of this transition was not well identified or understood by the ARIES development team at the time. Because the underlying data have a direct effect on each phase of the development process (decision model, data model, and system design), the data resource became a critical factor to the success and completion of the project.

A well thought out quality migration plan for source data can ease the labor of application development. [Ref. 20:p. 1] As a crucial piece that integrates the application, the migration plan is responsible for transforming, transporting, and scrubbing the data. This requires substantial prior planning.

### **1. Designate Migration Team**

The most important action to be taken in developing a data migration plan is to *identify the group of individuals that will be responsible and take ownership for the data portion of the project.* A team of individuals should be named early in the project development that will be responsible for making the required data available to the application. These team members should consist of both business and technical individuals from the application agency.

Members of the migration team will be responsible for gathering source files, analyzing those files, and maintaining the meta-data/business information about each file and its elements. This information about the source files is important to the process of mapping data elements to the individual decision measures. The understanding that can be gained by documenting the business knowledge underlying each data element and each data file will allow the development team to implement the business rules with a minimum of disruption. As each source file is analyzed and identified for use in the application, a process should be initiated for maintaining the business information about each file. This should include at a minimum the information that can be documented on

the Source File Documentation Form in Appendix D. The process of documenting the source files should be the responsibility of the business members of the migration team who are intimate with the source files.

The migration team will be responsible for the structure and architecture of the centralized data resource file that is based on the intended use of the data in the application and the structure of the source files. A migration strategy will be developed that includes determining a method of migration to maximize the information available to the decision support tool.

A migration strategy is achieved by determining what type of information will be maintained and at what level of detail. The term *granularity* is used in data warehousing to identify the level of detail that is contained in the data warehouse. The granularity is determined by either maintaining detailed data elements as they are found in the source files or by summarizing those elements to reduce the granularity. As discussed in earlier chapters summarizing or aggregating data in the data resource file can improve dramatically the performance of certain queries. Determining this strategy early in the project life cycle provides the system design team the advantage of a stable data source for a most of the overall project.

Once the design of the decision model is complete, the migration team can begin to assist in the development of the data model by assisting in the mapping of data elements to decision factors and in developing the business rules. The logic for the eventual data extraction process the team will use comes directly from these business rules.

## **2. Determine Extraction Logic and Generate Extraction Routine**

By being involved in the mapping of the decision measures to actual data elements, the migration team can begin the process of designing the logic comprising the data extraction process. The migration team will be required to identify the data elements required in order to reduce the amount of data that is migrated to the data resource file.

The logic involved in the extraction process can implement criteria that are in effect for any of the decision measures. An example of this logic would be filtering only the active reservists from the transaction files by identifying a specific code that is used to mark each active reservist.

The business rules developed in the data model can be used to identify common elements of aggregation and structure that are required in the data resource file. Extraction routines or queries can be developed to retrieve, transform, and summarize data into a format that will optimize the usability of the data resource file by the application. The extraction routines can also be used to determine and set the indexed fields of the data resource that allow the application's querying tool to access the stored data in an efficient manner.

Data scrubbing and cleansing for common data errors should also be implemented in the extraction routines. The data cleansing and conditioning rules are developed from the data quality assessments discussed in the next section.

### **3. Quality Assured Data**

As discussed in Chapters III and IV, data quality is an important issue that must be investigated very early in a project. The migration team will be directly involved in assessing the quality level of data available as well as that required by the proposed application. This portion of the migration process can do more to ensure the success of the project than any other. The first and most important task that must be undertaken pertaining to data quality is the Data Quality Baseline Assessment. Which will be used to identify problems with accuracy and inconsistencies of the source data that can be integrated into the data preprocessing routines. The baseline will also provide the customer and the development team with a deeper understanding of the data that are intended for use in the project.

In order for the migration team to be able to perform an analysis, they must first develop a set of quality metrics. This metrics should include, at a minimum, an

evaluation of the source files and their ability to meet the six characteristics of data quality outlined in Chapter III. By completing this initial assessment, the migration team will be able to identify potential problem areas and make a determination about the legitimacy of the data to support the decision process.

In conjunction with the customer, the migration team must determine and document the expectations for the quality of data. This expectation level should be based on the level of risk the customer is willing to accept if the resulting decision is wrong because of the underlying data. Meta-data and data quality metrics will be used to document these expectations and will be used on a continued basis to measure the performance of the data as it pertains to the expected level of data quality.

The data errors that result from comparing the source data with the data quality metrics will produce additional logic to be included in the extraction phase. This process, known as data conditioning or data cleansing will filter out potential data quality problems from the source files and thereby improve the data quality in the data resource file. This process will not necessarily correct the source of those problems, as quality problems can only be fully corrected in the source files. The migration team is responsible for identifying and correcting known data quality problems at their source wherever possible.

### **C. DECISION MODEL DEVELOPMENT**

A decision model is the foundation upon which the entire SDSS application will be built. In the SDSS concept, a known decision process is automated by accessing available database information. The reliability and trust placed on the outcome of the SDSS must be based on the quality of the underlying database information. Ken Orr states in his discussion of Data Quality and Systems Theory that the quality of data is a function of its use. [Ref. 14:p. 9] In the ARIES SDSS project many of the problems encountered during the development process were a result of the fact the data had not been used for the purpose of making site location decisions. Future applications should

attempt to leverage the data that is being used in current business processes to maximize the inherent quality factors arising from frequent use.

During the development of a decision model for a new application, care should be taken to document the associated data that is currently being used for decision-making. Identifying these sources will assist the migration team in their efforts to gather and present source data of the highest quality. In the case of the ARIES SDSS application, the source files were being used by a wide community of different users at USARC primarily for operational purposes as opposed to decision support applications. As a result, the data files were never obliged to meet the stringent quality standards required of DSS applications.

#### **D. DATA MODEL**

Problems with the underlying data will continue to be a problem in any application that attempts to leverage information stored in legacy databases. The goal in the future will be to minimize the effect these problems have on the development process. Listed below are three areas that can smooth the transition of data into the SDSS application and limit the impact on the application process.

##### **1. Data Standardization**

The elements of source data must be standardized when the legacy data files being used in an application are not constructed under the same set of business specifications. These inconsistencies in rules and definitions may lead to problems when the actual data are being interpreted out of the original context in which it was defined. Data standardization is achieved by logically identifying, grouping, and classifying data.

This lack of standards for the data attributes in existing applications can manifest in many ways. For example, data elements in different files may not label the same field in the same manner. Fields with the same name may not contain the same information because of differences in usage by different customers. Examples from the ARIES SDSS

project are the data fields that represent the Unit Identification Codes which were labeled differently in all of the following manners: UIC1, UIC, CURR\_UIC, OWN\_UIC. Furthermore, the UIC field was used to identify the same billet in different ways. The source file that lists the actual billets or jobs at a reserve facility was marked with the use of a parent UIC that identified the facility itself. The individuals actually assigned to a billet in the personnel file, on the other hand, are listed with a UIC that identifies the actual platoon to which the individual is assigned. Such inconsistencies make it difficult to verify one file against another. In this case it would be impossible to identify if there were a specific individual assigned for every billet in the billet structure file or whether the billet to which an individual is assigned is valid.

The purpose of data standards is to facilitate common use and understanding in identifying data characteristics. All parties involved in the project must be able to interpret the same information in exactly the same way. This will allow the development to be consistent and remove the need for each individual to have a deep understanding of each data file.

Data standardization must be conducted during the extraction process. The rules associated with the standardization specifications will be implemented in the extraction routines developed by the migration team. These specifications become part of the transformation process and the application can then be developed without concern about knowledge of the individual structure of source files.

## **2. Meta-data Documentation Process**

Understanding the information about the source databases, i.e., the meta-data, will allow the application to maximize the use of information in the data files. There were many times when development of the application stalled while the development team waited for insight about one or more source data files to be provided by the customer. Meta-data has two parts; (1) the detailed information about the data elements, their formats, length and so on, and (2) the business information and understanding about the

data file. The business information documents the rules involved in populating the data file, what the data file represents, and the criteria for each data element.

An important part of this documentation process is identifying the ownership of a source file and maintaining a knowledgeable point of contact for each file. These individuals will be responsible for documenting and maintaining the meta-data throughout the lifecycle of the project.

Michael Brackett describes the need for a meta-data warehouse that goes beyond the traditional data information storage repositories and provides a "personal help desk" for increasing the awareness and understanding of the data resource.[Ref. 3:p. 193] This concept would allow the user to access indexed information about the source file and therefore maximize his/her ability to identify what data are available. Using meta-data to the fullest extent possible would benefit the SDSS concept by allowing the decision-maker to tap the maximum amount of knowledge available in the source data files. The intent here is not create a meta-data documentation project but to provide the maximum amount of available information concerning the data files to the development team. This tedious and time consuming project will pay dividends in the quality of the output from the application.

### **3. Identify Spatial Aspects of Queries**

"Over 80% of business data have some spatial context such as a customer address, ZIP Code, or location." [Ref. 5:p. 1] Taking advantage of the spatial aspects of the underlying data can provide valuable information to the decision process as well as enhance the performance of the final product. Identifying the decision measures and their associated business rules that rely on a spatial aspect can be used to create a performance advantage.

During the ARIES SDSS project, complex queries that involved determining if one entry in a table existed in another table were found to have definite spatial aspects. The ARIES application realized a twenty-fold increase in query performance by simply



allowing the GIS application to conduct that portion of the query related to spatial parameters.

Using the advantages that a GIS system provides to localize data will allow the SDSS application to access larger quantities of data in a shorter amount of time. The ability of a SDSS application to access large quantities of data efficiently will allow incremental improvement of the underlying decision process. This spatial component may also allow the decision-maker to introduce new decision measures that can enhance the final outcome of the SDSS.

## **E. SYSTEM DESIGN**

Lessons learned from the system design portion of the ARIES project are discussed in a thesis being prepared concurrently by LT Peter Falk. As the principal designer of the UI application, his thesis provides a detailed discussion of the issues and challenges required to complete the ARIES SDSS prototype application.

## **F. FUTURE CONSIDERATIONS**

The concept of an SDSS application is still evolving. The ARIES prototype application has proved the viability of an asset that integrates a GIS system and DSS tools to leverage the knowledge maintained in legacy databases for decision-making purposes. Enhancements to be incorporated in methodologies used in future SDSS applications can be separated into the phases of the development process; Decision Model, Data Migration, Data Model, System Design, and Testing.

### **1. Decision Model**

Discussions of improvements to the decision model have been discussed fully in Reference 1 and Reference 2.

## **2. Data Migration**

The migration process implemented in the ARIES prototype only supports the application in its current configuration. The migration application, Aries Administrator, does not allow for additional data files or queries. This limitation will hinder the ability of the Administrator to support the ARIES application if any portion of the decision hierarchy is changed. Consideration should be given to adapting the Administrator application to allow the ability to add and remove files and queries from the system. The Administrator currently only allows for a complete extraction of every data file associated with the application, a time consuming event that is not necessary if every data file has not changed. A situation may arise where only one data file has changed; in this case, the Administrator application should be adapted to allow the user to conduct an intelligent update by choosing the data files that require updating. By documenting the update frequency of a data file in the meta-data (i.e., weekly, monthly, etc.), the Administrator could identify data files that have not been updated.

Based on the importance of the quality of data migrated to the data resource file, it will be important for future implementations of the SDSS methodology to have an automated method for determining and maintaining data quality. The migration process is the phase in which the rules associated with the quality metrics can be used to determine the quality of the underlying process. By automating this process, exceptions can be generated during the migration process that will identify known problems. The migration engine would be able to provide an estimation of the quality of data and determine whether that level is acceptable based on the expectations provided by the user.

## **3. Data Model**

Because the idea of a useful decision support tool involves the ability to be flexible as the decision process changes, an automated application such as ARIES must be able to adapt to that changing environment. There should be a system in place that will identify changes to any part of the application, (i.e., legacy data, decision model, data

quality requirements, etc.) and capture the effects of those changes on the application. For example, if a legacy database changes in any way the migration process should be adjusted to reflect those changes as well. Also, if a new decision measure is added, the data and meta-data to support that measure should be added to the data resource file. Flexibility of the application will be the key to its longevity. If the application cannot be updated easily as the business process or environment changes it will die a certain and swift death.

Data values that do not change over long periods of time and are used to support decision measures should be calculated only as those values change and not during each evaluation session. Values are calculated during each session for measures such as Distance to Recruit Station and Distance to ECS that do not change on a frequent basis. This calculation process could be moved from the evaluation process into the data migration process by computing a value that is pre-calculated for all possible sites. Early identification of values that change infrequently will reduce the overhead required in the system design portion of the project.

#### **4. System Design**

The ARIES SDSS prototype application instituted the use of an error value to help identify data that were missing or returned null (i.e., -999). This was effective for identifying a number of potential problems and provided valuable, albeit limited, feedback from the system directly to the user. However, this concept must be expanded to include other error codes for a more detailed feedback system. The error codes should be kept to a minimal list of highly useful codes. For example, other error codes should be used to identify a value of zero for a measure that should not be zero (i.e., Facility Age). Another example would be additional error values that can signify different types of problems resulting from the calculations, e.g., if Number Assigned returned for a unit is null or zero and a value is found for Losses or Transfers for the same unit, then an error exists in one of the files. Currently the ARIES application assigns a zero to the value of the calculation to avoid a division by zero. Adding an error value to identify

inconsistencies between data files would provide the user with a possible reason for the values of Loss Rate or Transfer Rate being zero. This type of error detection would require the application to have intelligent business rules that document the relationships between decision measures. This same process could also be used to identify possible problems with ratio values discussed in Chapter IV.

Other considerations were made to allow the user the ability to choose a default value for measures that returned with error values. The user could choose to use a previously determined value such as the mean. This would allow the decision model to include this measure in the evaluation of the site.

## **5. Testing**

The implementation and testing phase will continuously monitor and evaluate the results of the system and provide feedback to the application. The application must have the ability to monitor and evaluate the results and provide feedback to the system to improve the process. During the ARIES project, a problem like this was identified in the post implementation phase. Every value the system returned for Facility Condition, Measure #4, was the same value, GREEN. This resulted in each site receiving the maximum utility for that measure and effectively nullifying any benefits that measure provided to the decision model. A problem like this could be identified by observing the trends of the answers and having the testing module send a flag to the user that reports a problem with a decision measure. The application would provide some basic information to assist in diagnosing the problem. A full testing of the application requires a formal feedback mechanism that will allow problems to be documented as well as the solutions and corrections to be documented.

## **G. CHAPTER SUMMARY**

An inspection of the final ARIES SDSS prototype application can provide future implementations of the SDSS methodology with valuable information. The oversights,

problems, and mistakes discovered during the design and development of this “proof of concept” application has led to recommended changes to the development process. Among the changes are the addition of a DMP and minor refinements to the decision model, data model, and system design phases.

The need for a DMP was unfortunately realized too late in the ARIES project. Many of the stumbling blocks in the development could have been avoided had there been an integrated plan for the movement of data from its source to the DSS application. A detailed DMP will involve assigning responsible individuals to gather data, develop extraction logic, and generate extraction routines or queries. This process will provide a stable source of data as a foundation for the application. The migration team will also be responsible for evaluating the baseline quality of the data set and generating a plan to reach the desired level of quality in the final product.

Development of the decision model drives the entire SDSS development process and should be given the proper amount of attention. The failure of the ARIES project to identify sources receiving frequent use required the development team to spend valuable time validating and correcting the source data files. It is important to identify in the decision elicitation process the data elements that are currently being used in a system.

Because the data files used by the ARIES project were made up of a collection of large legacy data files from varying sources, there was a need for all members of the team to have the same understanding of what comprised these files. A system to maintain detailed information about the data files, i.e. meta-data, was required. Detailed documentation should accompany the transfer of a source file from the customer to the development team. Standardizing the common data elements in the data resource file provides the application with future flexibility. The ARIES project was able to leverage the spatial aspects inherent the underlying data that were directly associated with decision measures. Future SDSS implementations should make every attempt to harness the spatial aspects of the data.

As the SDSS methodology is used in future implementations, the need to add increased flexibility and feedback to the user will continue to enhance the usability of the final product. Future implementations will concentrate more attention on the quality of data and ability of the increased intelligence in the system to provide the user with highest quality decision support tool available.



## VI. CONCLUSION

### A. SUMMARY

Developing a Spatial Decision Support System (SDSS) for the Army Reserve TPU relocation decision problem provided insight into new methods to improve the development methodology for a SDSS. The Army Reserve Installation Evaluation System (ARIES) is the result of using this SDSS methodology to integrate a detailed decision model in an automated DSS. The system integrated two commercial software programs, Logical Decisions for Windows™ as a decision model solver and MapInfo™ as GIS mapping engine. A user interface (UI), created in Visual Basic™, served as an integration tool for retrieving data and passing information to between these components.

The system architecture developed for the ARIES project consisted of a decision model, a data model, and an integrating application. The decision model was developed under separate research and constituted the basis for gathering the required data to evaluate the readiness of an Army Reserve facility. The decision measures developed in the decision model generated a set of business rules that were mapped to actual data elements. Because of the complexity of the queries, the business rules and the quantity of data that was involved, the development team identified a need for a centralized data resource file. The ARIES data resource file used data warehousing techniques to conduct extractions from the many source data files, and was optimized for the ARIES decision process. A data preprocessing application was created to generate this data resource file. The ARIES Administrator is a Visual Basic™ application that acts as a migration engine to transform the source data into the structure required by the ARIES application.

Because of the spatial nature of the decision model involved, the ARIES data resource file used basic data warehousing techniques, such as aggregation and summarization, to take advantage of the spatial attributes of the source data. This spatially enabled data set is a special form of a data warehouse or data mart called a *spatial data warehouse*. The spatial aspects of the data were used in conjunction with the



GIS application to maximize query performance. The primary advantage of using geocoded (i.e., spatially identified) information in the queries was a significant increase in performance for the ARIES application.

Through the use of a spatial data warehouse, the SDSS is buttressed with a stable data source engineered to provide the underlying decision model with the highest quality data in a timely manner. The integration of a data migration plan (DMP) in the system development process ensures that the data resource generated for the application allows the SDSS application to generate meaningful outcomes.

## **B. CONTRIBUTIONS**

This research implemented the theoretical SDSS methodology by creating an integrated application in support of complex site location decisions. As a proof of concept application, ARIES demonstrates the ability to integrate a GIS mapping engine and a decision model solver in a seamless and flexible environment that allows users to leverage operational legacy database information for decision-making purposes. At an applied level, this research identified additional requirements necessary during the development process to provide SDSS applications with stable and accurate data sources of acceptable quality. These additional requirements involved the development of a data migration plan (DMP) and the implementation of a data quality assessment plan.

### **1. General Contributions**

In addition to the specific benefits afforded to USARC, this project identified enhancements to existing SDSS methodology development to ensure data quality. Most important is the requirement to transport and transform the underlying data into a format that allows the SDSS application to access that data in the most efficient manner. The DMP that is outlined in this paper provides the basis for a data resource to instantiate a decision model in a fashion that improves performance and assures a confidence in of the outcome.

Data quality was identified as a limiting factor of the SDSS application too fully analyze the site evaluation as well as provide an outcome that is credible to the user. This identified the need to incorporate the evaluation and assessment of source file data quality as a continuing effort throughout the development process. An important element in correcting and maintaining an expected level of quality for data is the assignment of individual responsibility for identifying and correcting the inadequacies of the source data files.

## **2. Specific Contributions to USARC**

The primary benefit of the ARIES project is the use of a very powerful decision tool to provide the decision-maker with detailed information previously not available. By implementing the detailed and complex queries that provides values for the ARIES decision measures, USARC has benefited by being able to analyze this information. The ARIES application goes far beyond mere data retrieval, allowing the decision-maker to manipulate the results of these complex queries in a highly flexible and fully functional decision environment.

This research showed by detailed analysis that 14 of the 20 decision measures that have been automated will return a valid value more than 90% of the time. Further data quality analysis would provide the USARC Readiness team with the assurance that the ARIES application is basing its outcome on data that are accurate, consistent, complete, timely, and unique, as well as valid.

Through implementation of the Administrator, USARC has benefited from spatial data warehousing techniques to improve performance of the system by centralizing the data elements required for the TPU relocation decision problem. The Administrator provides a stable data set to the ARIES application by using queries that can be repeated time after time as the source files change. The Administrator also provides an automated data quality filter that facilitates data cleansing of the source data sets.

Even without an implementable SDSS application, USARC has received the benefit of an in depth look at the data files they are using in their everyday decision-making. It has forced the group of experts to verify and validate the assumptions they may have made concerning the site location decision problem.

The real value of this research may lie in the basis it provides for future SDSS applications to increase access to decision information directly from legacy data sources. Developing a strategy to provide SDSS with high quality data creates a foundation for a much higher probability of successful implementation of decision-based systems.

## APPENDIX A. DECISION MODEL MEASURES

This appendix contains detailed information about each decision measure that was automated in the ARIES SDSS prototype application. The information includes a description of each measure, the business rule used to calculate the associated value, base units, source files, associated ACROPOLIS tables, query or queries involved in the calculation, a description of the yield curve, and a graph of the yield curves. "ACROPOLIS" is the file name for the ARIES data resource file.

The term "in the area" in this Appendix is defined as being within a 50-mile radius of the moving unit or proposed facility.

### Index

Measure 1. Facility Backlogged Maintenance .....	79
Measure 2. Facility Operating Costs .....	81
Measure 3. Facility Age .....	83
Measure 4. Facility Condition .....	85
Measure 5. Facility Ownership .....	87
Measure 6. Competition .....	89
Measure 7. Average Area Drill Attendance .....	91
Measure 8. Area Loss Rate .....	93
Measure 9. Area Transfer Rate .....	95
Measure 10. Area Average Manning .....	97
Measure 11. Distance to Nearest Recruit Station .....	99
Measure 12. Available Transfers from Closing Units .....	101
Measure 13. IRR Available .....	103
Measure 14. Recruit Market .....	105
Measure 15. Reassignments .....	107
Measure 16. Distance to Area Maintenance Support Activity .....	109
Measure 17. Distance to Nearest Equipment Concentration Site .....	111
Measure 18. Facility Weekends Used .....	113
Measure 19. Available MOS from Closing Units .....	115
Measure 20. Available MOS IRR .....	119

THIS PAGE LEFT INTENTIONALLY BLANK

## Measure 1. Facility Backlogged Maintenance

**Definition:** Facility Backlogged Maintenance provides the total dollar value of backlogged maintenance. This provides an indication of the initial investment required to correct the significant maintenance problems with a proposed facility.

**Calculation:** The Backlogged Maintenance value is based upon the sum values for maintenance actions documented for each facility in the "CWE\_TOTAL" field of the RPINFODT file. The summation is done during the data extraction phase.

Maint\_Cost[Sum of outstanding maintenance actions for a facility]

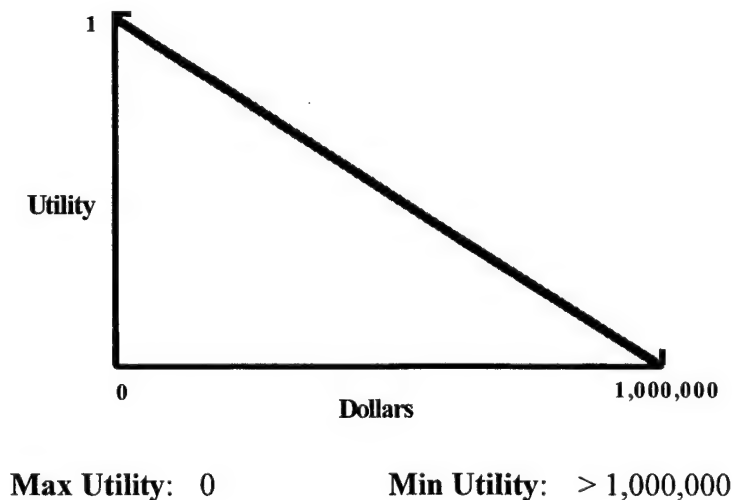
**Units:** Dollars

**Source File:** RPINFODT

**ACROPOLIS Table(s):** RPINFODT\_

**Query:** Maint\_Cost  
SELECT MAINT\_COST  
FROM RPINFODT\_  
WHERE RPINFODT\_.FACID = ProposedFacility.FAC\_ID

**Yield Curve:** A linear relationship is assumed between the backlogged maintenance costs and utility. Every dollar required or saved in this category is expected to have equal utility to a relocating unit.



THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 2. Facility Operating Costs

**Definition:** Facility Operating Costs provide an indication of the financial resources that are required to maintain the facility in a serviceable condition. This includes both utilities and minor maintenance costs.

**Calculation:** Operating Costs are extracted from the "COST\_PR\_SF" field of the FPS file.

COST\_PR\_SF[Retrieve the Cost per Square Foot for a facility]

**Units:** Dollars per square foot per month

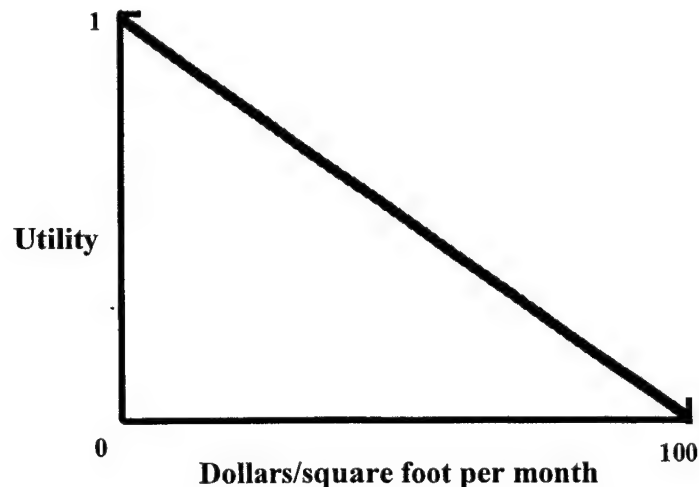
**Source File:** FPS

**ACROPOLIS Table(s):** FPS\_

**Query:**

```
COST_PR_SF
SELECT COST_PR_SF
FROM FPS_
WHERE FPS_.FACID = ProposedFacility.FAC_ID
```

**Yield Curve:** A linear relationship is assumed between the operating costs and utility. Every dollar required or saved in this category is expected to have equal utility to a relocating unit.



Max Utility: 0

Min Utility: > 100



THIS PAGE INTENTIONALLY LEFT BLANK

### Measure 3. Facility Age

**Definition:** This measure indicates the age of the primary structure on the proposed relocation site. It is intended to reflect an assumed long term structural degradation with time.

**Calculation:** Facility age is calculated based upon the acquisition date found in the INTEREST file. The acquisition date is compared to the current date and the difference is determined in months.

DATE\_ACQ[Current Year - Date Acquired]

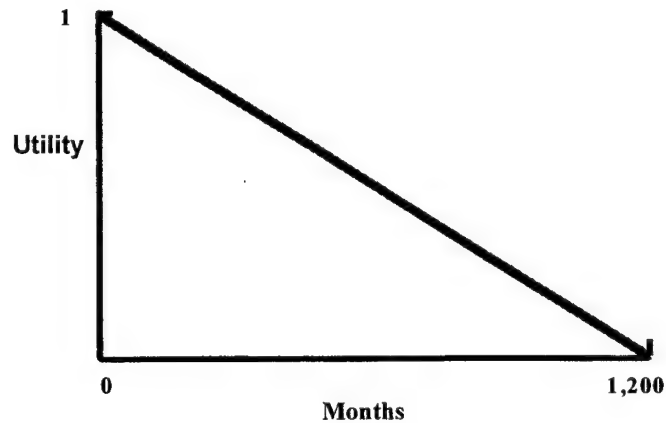
**Units:** Months

**Source File:** INTEREST

**ACROPOLIS Table(s):** INTEREST\_

**Query:** DATE\_ACQ  
SELECT DATE\_ACQ  
FROM INTEREST\_  
WHERE INTEREST\_.FACID = ProposedFacility.FAC\_ID

**Yield Curve:** A linear relationship is used between facility age and utility.



Max Utility: 0

Min Utility: > 1,200

THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 4. Facility Condition

**Definition:** Facility Condition is based upon a visual inspection of the structure and provides an indication of the serviceability of the primary structures.

**Calculation:** This measure is based upon the ISR part 1 rating entered in the "FAC\_COND" field of the FPS file.

FAC\_COND[Retrieve Facility Condition]

**Units:** No Units(Green, Amber, Red)

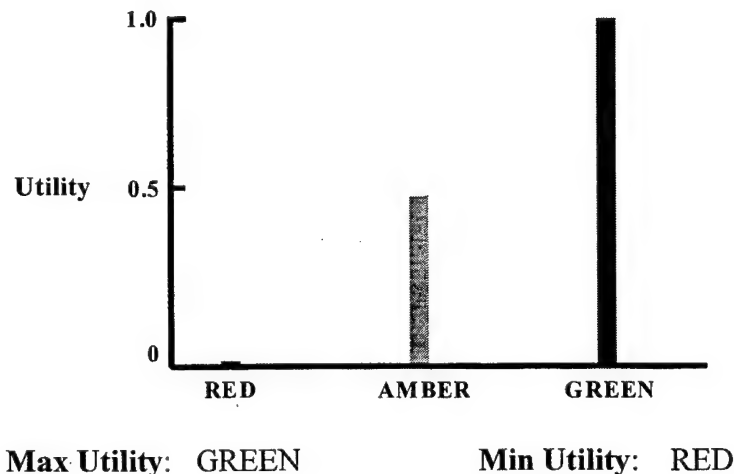
**Source File:** FPS

**ACROPOLIS Table(s):** FPS\_

**Query:**

```
FAC_COND
SELECT FAC_COND
FROM   FPS_
WHERE  FPS_.FACID = ProposedFacility.FAC_ID
```

**Yield Curve:** The utility of these three categories varies in discrete steps. A facility that is categorized as "green" is judged to be approximately twice as desirable as one that is assigned an "amber" rating.



THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 5. Facility Ownership

**Definition:** This measure indicates whether the facilities at a proposed relocation site are leased or owned.

**Calculation:** Facility Ownership is based upon the entry in the "GOVT\_OWN" field of the COMPLEX file.

GOVT\_OWN[Retrieve Ownership Status]

**Units:** No Units(Yes/No)

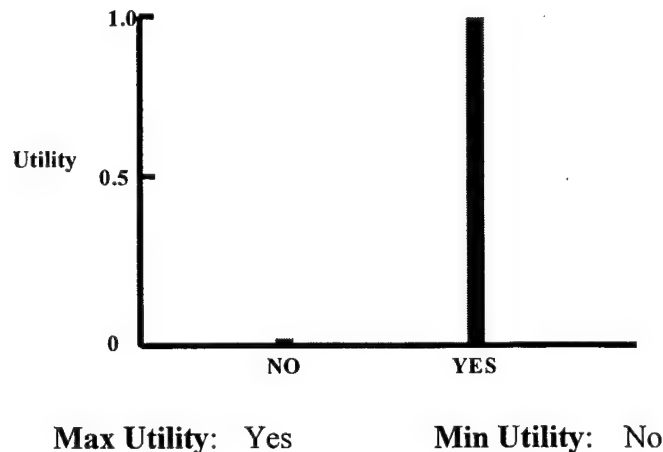
**Source Data:** COMPLEX

**ACROPOLIS Table(s):** COMPLEX\_

**Query:**

```
GOVT_OWN
SELECT GOVT_OWN
FROM COMPLEX_
WHERE COMPLEX_.FACID = ProposedFacility.FAC_ID
```

**Yield Curve:** Facilities that are owned by the government are preferred as relocation sites over those facilities that are leased. The owned sites are assigned the maximum utility value of 1.0, while leased sites are given a 0 utility score.



THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 6. Competition

**Definition:** This measure provides an indication of the level of competition for potential reservists. It considers only Army Reserve and Army National Guard units in the area of the relocation site.

**Calculation:** Competition is determined by the number of positions that must be filled by all other Army Reserve and Army National Guard (ARNG) units in the area of the proposed relocation site. For Army Reserve units, the number of required positions is determined by counting the number of records in the G19TRUE file associated with each UIC in the area. For ARNG units, the value is found in the "AUTH" field of the NGNON\_CL file.

$$\text{NO\_AUTH\_NG}[\text{Number Authorized National Guard}] + \text{NO\_REQD}[\text{Number Area Reservists Required}]$$

**Units:** Number of competing positions

**Source File:** COMMAND PLAN, G17, G19TRUE, GEOREF, NGNON\_CL

**ACROPOLIS Table(s):** CMDPLAN, G17Natl, G19Natl, VALID\_UIC

**Query:** Area-FACID List(MapInfo)  
SELECT FAC\_ID INTO TempFACID  
FROM GEOREF  
WHERE Object Within ObjAreaBuffer  
ORDER BY FAC\_ID  
(Note: ObjAreaBuffer is equal to 300 miles)

VALID\_UIC  
SELECT UIC, FAC\_ID, UnitName, City, State, Zip  
FROM G17Natl  
WHERE G17Natl.UIC = ANY (SELECT CMDPLAN.UIC  
FROM CMDPLAN)

Area-UIC List  
SELECT DISTINCT UIC INTO AREA\_UIC  
FROM VALID\_UIC  
WHERE VALID\_UIC.FAC\_ID = ANY (SELECT AREA\_FACID.FAC\_ID  
FROM AREA\_FACID)

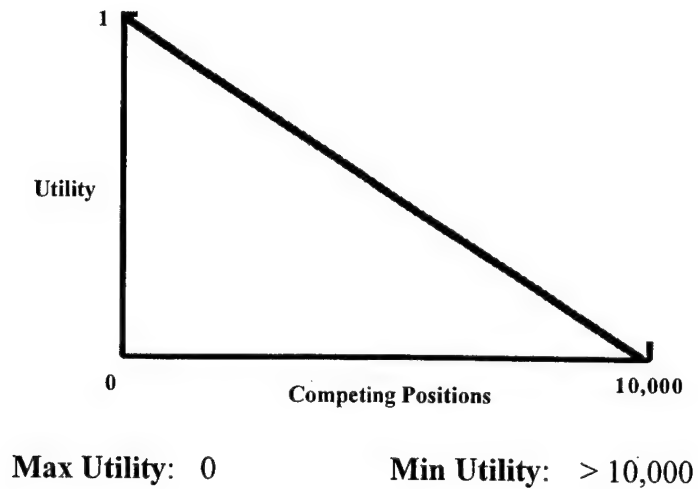
NO\_AUTH\_NG(MapInfo)  
SELECT \* INTO TempNGUnits  
FROM NON\_CLOS  
WHERE Obj Within ObjAreaBuffer



```
SELECT SUM(AUTH) "No_AUTH_NG" INTO Strength
FROM   TempNGUnits
```

```
NO_REQD
SELECT SUM(UIC_TOTAL) AS TOTAL_REQD
FROM   G19NatI
WHERE  G19NatI.UIC = ANY (SELECT AREA_UIC.UIC
                          FROM AREA_UIC)
```

**Yield Curve:** A linear relationship exists between the number of competing positions from other units and the utility of a relocation site. The level of no site utility in this measure begins at 10,000 positions which is above the maximum value expected.



## Measure 7. Average Area Drill Attendance

**Definition:** This measure indicates the fraction of reservists with satisfactory drill attendance for all existing units in the area of the proposed relocation site. Areas with a high fraction of satisfactory drill attendance are preferred relocation sites because units relocated to that area are assumed to perform similarly in drill attendance.

**Calculation:** This measure considers the last four quarters of data contained in the FINANCE file. After initial screening, the number of reservist with 21 or more drill periods for the year is divided by the total number of people who meet the screening.

$$\frac{\text{DRILL\_SAT [Number of reservists with > 21 drill periods in a year]}}{\text{DRILL\_TOTAL [Number of reservists required to drill]}}$$

**Units:** Ratio

**Source File:** COMMAND PLAN, FINANCE, G17, G19TRUE, GEOREF

**ACROPOLIS Table(s):** CMDPLAN, FINANCE\_, FINANCE\_QTR, G17Natl, G19Natl, VALID\_UIC

**Query:** Area-FACID List(MapInfo)  
SELECT FAC\_ID INTO TempFACID  
FROM GEOREF  
WHERE Object Within ObjAreaBuffer  
ORDER BY FAC\_ID  
(Note: ObjAreaBuffer is equal to 300 miles)

VALID\_UIC  
SELECT UIC, FAC\_ID, UnitName, City, State, Zip  
FROM G17Natl  
WHERE G17Natl.UIC = ANY (SELECT CMDPLAN.UIC  
FROM CMDPLAN)

Area-UIC List  
SELECT DISTINCT UIC INTO AREA\_UIC  
FROM VALID\_UIC  
WHERE VALID\_UIC.FAC\_ID = ANY (SELECT AREA\_FACID.FAC\_ID  
FROM AREA\_FACID)

```

FINANCE_CY
SELECT UIC, COUNT(UIC) AS UIC_TOTAL INTO FINANCE_CY
FROM FINANCE_QTR
WHERE (Select Case)
      Case 1st Qtr FY
      (UTA1Q1PF + UTA2Q1PF + UTA3Q1PF + UTA4Q1PF) > 20
      Case 2nd Qtr FY
      (UTA2Q1PF + UTA3Q1PF + UTA4Q1PF + UTA1QCFY) > 20
      Case 3rd Qtr FY
      (UTA3Q1PF + UTA4Q1PF + UTA1QCFY + UTA2QCFY) > 20
      Case 4th Qtr FY
      (UTA4Q1PF + UTA1QCFY + UTA2QCFY + UTA3QCFY) > 20
GROUP BY UIC
ORDER BY UIC

```

```

DRILL-SAT
SELECT SUM(UIC_TOTAL) AS TOTAL_SAT
FROM FINANCE_CY
WHERE FINANCE_CY.UIC = ANY (SELECT AREA_UIC.UIC
                             FROM AREA_UIC)

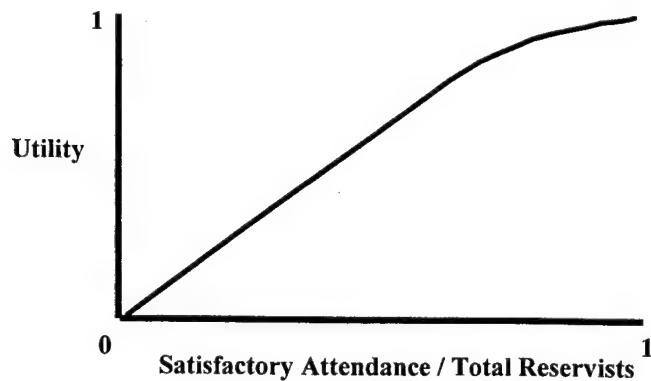
```

```

DRILL-TOTAL
SELECT SUM(UIC_TOTAL) AS DRILL_TOTAL
FROM FINANCE_
WHERE FINANCE_.UIC = ANY (SELECT AREA_UIC.UIC
                           FROM AREA_UIC.UIC)

```

**Yield Curve:** The utility of the average drill attendance rate increases linearly between the values of 0 and 0.6. Above that point, increases in the attendance rate result in diminishing returns. Values above 0.6 become increasingly uncommon.



Max Utility: 1.0

Min Utility: 0.0

## Measure 8. Area Loss Rate

**Definition:** This measure indicates the fraction of reservists who left the reserves in the previous fiscal year, for all existing units in the area of the proposed relocation site. Areas with a low loss rate are preferred relocation sites because units relocated to that area will also experience low loss rates.

**Calculation:** The number of losses to units in the area in the previous fiscal year is divided by the number of reservists currently assigned to these units. Losses are identified through the transfer mnemonic field (TRMN="LOSS") of the FyxxLOSS file. The number of assigned reservists is determined by counting all of the personnel records in the G18CWE file associated with each UIC in the area.

$$\frac{\text{NO\_LOSS[Total Number of Losses in the last year]}}{\text{NO\_ASSN[Total Number Reservists Assigned]}}$$

**Units:** Ratio

**Source File:** COMMAND PLAN, FYxxLOSS, G17, G18CWE, GEOREF

**ACROPOLIS Table(s):** CMDPLAN, FYxxLOSS, G17Natl, G18Natl\_UIC, VALID\_UIC

**Query:** Area-FACID List(MapInfo)  
SELECT FAC\_ID INTO TempFACID  
FROM GEOREF  
WHERE Object Within ObjAreaBuffer  
ORDER BY FAC\_ID  
(Note: ObjAreaBuffer is equal to 300 miles)

VALID\_UIC  
SELECT UIC, FAC\_ID, UnitName, City, State, Zip  
FROM G17Natl  
WHERE G17Natl.UIC = ANY (SELECT CMDPLAN.UIC  
FROM CMDPLAN)

Area-UIC List  
SELECT DISTINCT UIC INTO AREA\_UIC  
FROM VALID\_UIC  
WHERE VALID\_UIC.FAC\_ID = ANY (SELECT AREA\_FACID.FAC\_ID  
FROM AREA\_FACID)

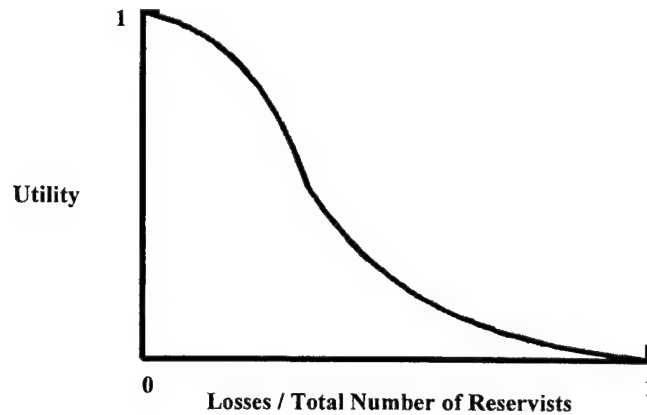
NO\_ASSN  
SELECT SUM(UIC\_TOTAL) AS TOTAL\_ASSN  
FROM G18Natl\_UIC  
WHERE G18Natl\_UIC.UIC = ANY (SELECT AREA\_UIC.UIC  
FROM AREA\_UIC)

```

NO_LOSS
SELECT SUM(UIC_TOTAL) AS TOTAL_LOSS
FROM   FYxxLOSS
WHERE  FYxxLOSS.UIC = ANY (SELECT AREA_UIC.UIC
                           FROM AREA_UIC)

```

**Yield Curve:** This function includes both concave and convex regions. The inflection point occurs at a loss rate of .33 and a utility of 0.5. Based on experience, a loss rate of one third per year was considered to be typical. Any loss rate below this value has relatively high utility, whereas loss rates above the inflection point quickly approach a utility of zero.



Max Utility: 0

Min Utility: 1

## Measure 9. Area Transfer Rate

**Definition:** This measure indicates the fraction of reservists who transferred to different units in the previous fiscal year for all existing units in the area of the proposed relocation site. Areas with a low transfer rate are preferred relocation sites because units relocated to that area will also experience low transfer rates.

**Calculation:** The number of transfers in the previous fiscal year is divided by the number of reservists currently assigned to the unit. Transfers are identified through the transfer mnemonic field (TRMN="TRFD") of the FyxxLOSS file. The number of assigned reservists is determined by counting all of the personnel records in the G18CWE file associated with each UIC.

$$\frac{\text{NO\_XFER[Total Number of Transfers in the last year]}}{\text{NO\_ASSN[Total Number Reservists Assigned]}}$$

**Units:** Ratio

**Source File:** COMMAND PLAN, FYxxLOSS, G17, G18CWE, GEOREF

**ACROPOLIS Table(s):** CMDPLAN, G17Natl, G18Natl\_UIC, FYxxXFER, VALID\_UIC

**Query:** Area-FACID List(MapInfo)  
SELECT FAC\_ID INTO TempFACID  
FROM GEOREF  
WHERE Object Within ObjAreaBuffer  
ORDER BY FAC\_ID  
(Note: ObjAreaBuffer is equal to 300 miles)

VALID\_UIC  
SELECT UIC, FAC\_ID, UnitName, City, State, Zip  
FROM G17Natl  
WHERE G17Natl.UIC = ANY (SELECT CMDPLAN.UIC  
FROM CMDPLAN)

Area-UIC List  
SELECT DISTINCT UIC INTO AREA\_UIC  
FROM VALID\_UIC  
WHERE VALID\_UIC.FAC\_ID = ANY (SELECT AREA\_FACID.FAC\_ID  
FROM AREA\_FACID)

```

NO_ASSN
SELECT SUM(UIC_TOTAL) AS TOTAL_ASSN
FROM    G18Natl_UIC
WHERE   G18Natl_UIC.UIC = ANY (SELECT AREA_UIC.UIC
                                FROM AREA_UIC)

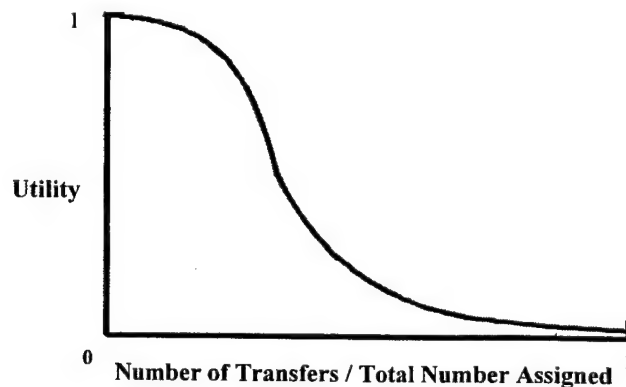
```

```

NO_XFER
SELECT SUM(UIC_TOTAL) AS TOTAL_XFER
FROM    FYxxXFER
WHERE   FYxxXFER.UIC = ANY (SELECT AREA_UIC.UIC
                             FROM AREA_UIC)

```

**Yield Curve:** This function includes both concave and convex regions. The inflection point occurs at a loss rate of .33 and a utility of 0.5. Based on experience, a transfer rate of one third per year was considered to be typical. Any loss rate below this value has relatively high utility (close to 1.0), whereas loss rates above the inflection point quickly approach a utility of zero.



Max Utility: 0

Min Utility: 1

## Measure 10. Area Average Manning

**Definition:** This measure indicates the ability to fill the required positions. An average value is determined for all existing units in the area of the proposed relocation site. Areas with high average manning levels are preferred relocation sites because units relocated to that area will also experience high manning levels.

**Calculation:** The number of reservists assigned to area units (based upon the number of personnel records in G18CWE file associated with each UIC) is divided by the number of required positions (based upon the number of positions in the G19TRUE file associated with each UIC). An average is calculated for all UIC's in the area of the proposed site.

$$\frac{\text{NO\_ASSN[Total Number Reservists Assigned]}}{\text{NO\_REQD[Number Area Reservists Required]}}$$

**Units:** Ratio

**Source File:** COMMAND PLAN, G17, G18CWE, G19TRUE, GEOREF

**ACROPOLIS Table(s):** CMDPLAN, G17Natl, G18Natl\_UIC, G19Natl, VALID\_UIC

**Query:** Area-FACID List(MapInfo)  
SELECT FAC\_ID INTO TempFACID  
FROM GEOREF  
WHERE Object Within ObjAreaBuffer  
ORDER BY FAC\_ID  
(Note: ObjAreaBuffer is equal to 300 miles)

VALID\_UIC  
SELECT UIC, FAC\_ID, UnitName, City, State, Zip  
FROM G17Natl  
WHERE G17Natl.UIC = ANY (SELECT CMDPLAN.UIC  
FROM CMDPLAN)

Area-UIC List  
SELECT DISTINCT UIC INTO AREA\_UIC  
FROM VALID\_UIC  
WHERE VALID\_UIC.FAC\_ID = ANY (SELECT AREA\_FACID.FAC\_ID  
FROM AREA\_FACID)



```

NO_ASSN
SELECT SUM(UIC_TOTAL) AS TOTAL_ASSN
FROM   G18Natl_UIC
WHERE  G18Natl_UIC.UIC = ANY (SELECT AREA_UIC.UIC
                               FROM AREA_UIC)

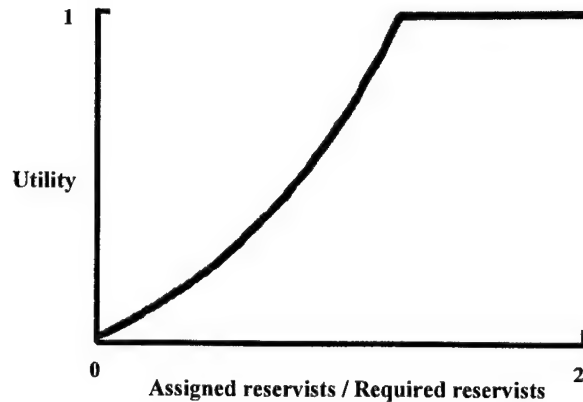
```

```

NO_REQD
SELECT SUM(UIC_TOTAL) AS TOTAL_REQD
FROM   G19Natl
WHERE  G19Natl.UIC = ANY (SELECT AREA_UIC.UIC
                          FROM AREA_UIC)

```

**Yield Curve:** It is desirable that area units be able to exceed their minimum manning requirements. All manning levels above 125% are considered to have maximum utility. Manning levels below this value drop off quickly in terms of utility.



**Max Utility:** 1.25

**Min Utility:** 0

## Measure 11. Distance to Nearest Recruit Station

**Definition:** Distance to the nearest Recruiting Station provides one indication of recruiter effectiveness.

**Calculation:** The straight-line distance from the proposed site to the closest recruiting station is calculated using a geocoded version of the RZA file.

DIST\_RZA[Determine distance to nearest Recruit Station]

**Units:** Miles

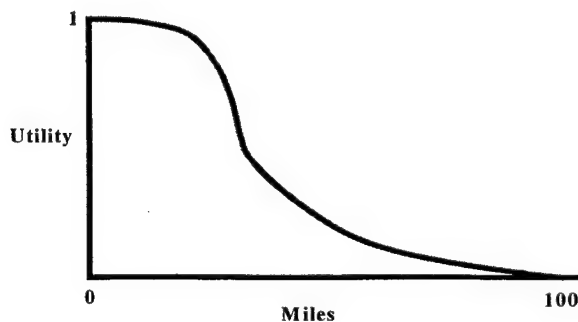
**Source Data:** RZA

**ACROPOLIS Table(s):** NONE

**Query:** DIST\_RZA(MapInfo)  
SELECT \*  
FROM RZA  
WHERE Obj Withing ObjDistanceBuffer into TempRZA  
(Note: ObjDistanceBuffer is equal to 300 miles)

```
SELECT Distance((CentroidX(Obj), CentroidY(Obj), FacIDLat, FacIDLong, "mi")
FROM TempRZA
ORDER BY Distance INTO TempRZA.Dist
```

**Yield Curve:** The effectiveness of a recruiting station in filling positions at a reserve unit is fairly high if the two are within a half hour drive of each other. It is assumed that recruiters are most effective in the area close to their recruiting station and that reserve recruits must be located near the unit with which they will serve. A distance of 30 miles is assigned an average utility of 0.5. A small change in distance results in less change in desirability when the distance is very small or very large than it does when the distance is around 30 miles.



**Max Utility:** 0

**Min Utility:** > 100

THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 12. Available Transfers from Closing Units

**Definition:** This value indicates the total number of personnel assigned to closing units within 50 miles of the proposed site.

**Calculation:** A list of Unit Identification Codes (UIC's) is created which contains only those units scheduled to close within 18 months. These units are identified by an entry of 5B in the "Tier" field of the G17 file. The number of potential transfers from closing units is calculated by summing the number of records in the G18CWE database for the closing units which are located in the area of the proposed relocation site.

TOTAL\_AVAIL[Total Number of Available Reservists from Area Closing Units]

**Units:** Ratio

**Source File:** COMMAND PLAN, G17, G18CWE, GEOREF, US\_ZIPS(MapInfo)

**ACROPOLIS Table(s):** CMDPLAN, G17NatI, VALID\_UIC

**Query:** Area-FACID List(MapInfo)  
SELECT FAC\_ID INTO TempFACID  
FROM GEOREF  
WHERE Obj Within objAreaBuffer  
ORDER BY FAC\_ID  
(Note: objAreaBuffer is equal to 300 miles)

VALID\_UIC  
SELECT UIC, FAC\_ID, UnitName, City, State, Zip  
FROM G17NatI  
WHERE G17NatI.UIC = ANY (SELECT CMDPLAN.UIC  
FROM CMDPLAN)

Area-UIC List  
SELECT DISTINCT UIC INTO AREA\_UIC  
FROM VALID\_UIC  
WHERE VALID\_UIC.FAC\_ID = ANY (SELECT AREA\_FACID.FAC\_ID  
FROM AREA\_FACID)

AREA\_CLOS\_UIC  
SELECT UIC  
FROM G17NatI  
WHERE G17NatI.TIER = "5B"  
AND G17NatI.UIC = ANY (SELECT AREA\_UIC.UIC  
FROM AREA\_UIC)

```

AREA_ZIPCODE(MapInfo)
SELECT ZIP_CODE
FROM US_ZIPS
WHERE Obj Within objAreaBuffer
ORDER BY ZIP_CODE

```

```

Area_G18_ZIP(MapInfo)
SELECT DISTINCT UIC, ZIPCODE, COUNT(UIC) AS UIC_TOTAL
FROM G18CWE
GROUP BY UIC, ZIPCODE
ORDER BY UIC, ZIPCODE

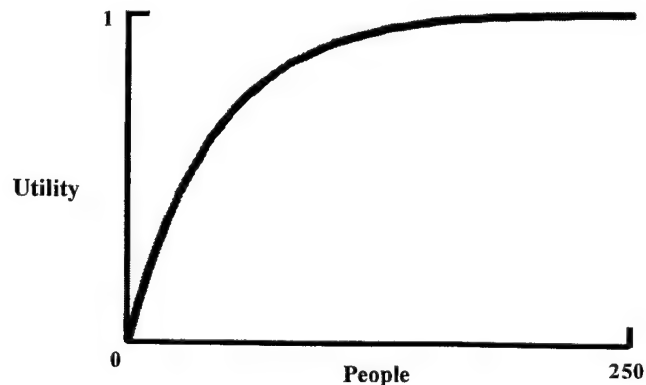
```

```

TOTAL_AVAIL
SELECT SUM(UIC_TOTAL) AS TOTAL_AVAIL
FROM Area_G18_ZIP
WHERE Area_G18_ZIP.UIC = ANY (SELECT AREA_CLOS_UIC.UIC
                                FROM AREA_CLOS_UIC)
AND Area_G18_ZIP.ZIPCODE = ANY (SELECT AREA_ZIPCODE.ZIP
                                FROM AREA_ZIPCODE)

```

**Yield Curve:** The shape of this function assumes diminishing returns in the number of transfers available. Experience suggests that for an average unit of 100 people, approximately half have prior reserve experience and that approximately half of the people in a closing unit will be able to transfer their skills directly to a new unit. The value of the first 100 reservists increases at a nearly linear rate because they provide preferred fills for approximately 50 of the positions of the moving unit. A value of 100 personnel is assigned a utility of 0.9. The incremental value added by each additional person over 100 continues to drop until no marginal gain is expected over 500.



**Max Utility:** > 250

**Min Utility:** 0

### Measure 13. IRR Available

**Definition:** Individual Ready Reserve (IRR) Available is the number of IRR members living in the area of the proposed relocation site. This is a measure of the size of the prior service market.

**Calculation:** A geographical query returns the total number of IRR members living within a specified distance of the proposed relocation site. This process requires a geocoded version of the IRR file.

TOTAL\_IRR[Total Number of Available IRR from the Area]

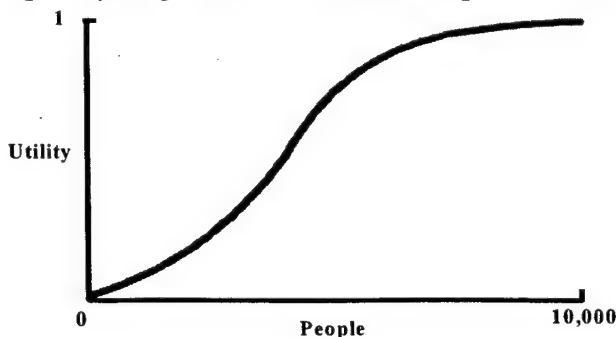
**Units:** People

**Source File:** IRR

**Query:** AreaIRR(MapInfo)  
SELECT ZIPC "ZIP", LEFT\$(PMOS, 3) "MOS"  
FROM IRR  
WHERE Obj Within objAreaBuffer  
AND ZIPC <> "" AND PMOS <> ""  
ORDER BY ZIPC  
(Note: objAreaBuffer is equal to 300 miles)

TOTAL\_IRR  
SELECT COUNT(\*) AS TOTAL\_IRR  
FROM AreaIRR

**Yield Curve:** For a typical unit of 100 people, it is assumed that approximately 40 positions could best be filled by IRR members. The recruiting rate for the IRR is approximately 1 percent, so an area that offers 4,000 IRR members is assigned an average utility of 0.5. Above this point, there are diminishing returns. The market begins to exceed the personnel demand of a moving unit and limited recruiting efforts become marginally less effective. The utility of smaller numbers quickly drops off because of the importance of this source of recruits.



Max Utility: > 10,000

Min Utility: 0

THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 14. Recruit Market

- Definition:** The Recruit Market measure estimates the total number of males who:
1. live in the area of the proposed relocation site
  2. Would score in the top half on the Armed Forces Qualification Test (AFQT)
  3. Fall into the desired age group (17 - 29 years old)
- Calculation:** This measure sums the entries for all mental categories 1 through 3A, and all ethnic groups for the zip codes of interest in the Qualified Military Available (QMA) file. The version of QMA used contains only the estimates for males within the age range of 17 to 29.

TOTAL\_MARKET[Total Non-Prior Service Personnel from the Area]

**Units:** People

**Source File:** QMA, US\_ZIPS(MapInfo)

**ACROPOLIS Table(s):** NONE

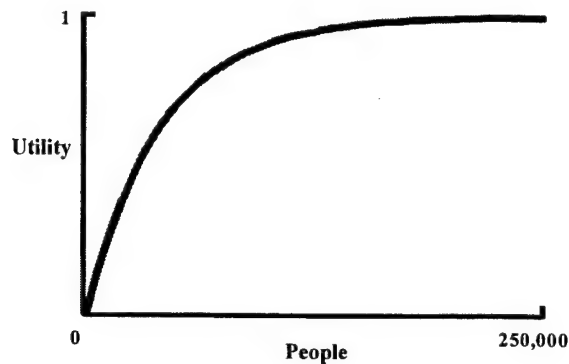
**Query:** QMA(MapInfo)  
SELECT LEFT\$(ZIP, 5) "ZIPCODE", MWCAT12, MWCAT3A, MBCAT12,  
MBCAT3A, MHCAT12, MHCAT3A  
FROM QMA  
WHERE Obj Within objAreaBuffer  
ORDER BY ZIP  
(Note: objAreaBuffer is equal to 300 miles)

AREA\_ZIPCODE(MapInfo)  
SELECT ZIP\_CODE  
FROM US\_ZIPS  
WHERE Obj Within objAreaBuffer  
ORDER BY ZIP\_CODE

TOTAL\_MARKET  
SELECT SUM(MWCAT12+MWCAT3A+MBCAT12+MBCAT3A+  
MHCAT12+MHCAT3A) AS TOTAL\_MARKET  
FROM QMA  
WHERE QMA.ZIP = ANY (SELCET AREA\_ZIPCODE.ZIP  
FROM AREA\_ZIPCODE)



**Yield Curve:** Approximately half of a typical unit of 100 reservists is filled by recruits with no prior service. Assuming a recruit rate of 0.25 percent, there must be at least 20,000 people in the area of the proposed relocation site who meet all of the requirements stated above. This value is assigned a typical utility of 0.5. As the number increases, there are diminishing returns. The market begins to exceed the personnel demand of a moving unit and limited recruiting efforts become marginally less effective.



**Max Utility:** > 250,000

**Min Utility:** 0

## Measure 15. Reassignments

**Definition:** The Reassignments measure indicates the fraction of the reservists assigned to the moving unit who currently live within a specified distance (50 miles) of the proposed relocation site

**Calculation:** This measure is calculated by first determining all zip codes that lie within a specified distance of the proposed relocation site (based upon zip code centroid) and then identifying all reservists who both live within one of the identified zip codes (based upon the "ZIP" field of the G18CWE file) and are assigned to the moving unit (based upon the "UIC" field of the G18CWE file). Then the number available reassignments is divided by the total number of reservists assigned to the moving unit.

$$\frac{\text{TOTAL\_RESERVISTS[Total Number of Available Reservists from the Moving Unit]}}{\text{UIC\_TOTAL[Total Number of Reservists Assigned Moving Unit]}}$$

**Units:** Ratio

**Source File:** G18CWE, US\_ZIPS(MapInfo)

**ACROPOLIS Table(s):** G18Nat1

**Query:** AREA\_ZIPCODE(MapInfo)  
SELECT ZIP\_CODE  
FROM US\_ZIPS  
WHERE Obj Within objAreaBuffer  
ORDER BY ZIP\_CODE

G18(MapInfo)  
SELECT UIC, LEFT\$(ZIP,5) "ZIPCODE", PRI "MOS"  
FROM G18CWE  
WHERE Obj Within objG18Buffer AND PRI <> ""  
ORDER BY UIC, ZIP  
INTO G18

Area\_G18\_ZIP  
SELECT DISTINCT UIC, ZIPCODE, COUNT(UIC) AS UIC\_TOTAL  
FROM G18  
GROUP BY UIC, ZIPCODE  
ORDER BY UIC, ZIPCODE

```

TOTAL_RESERVISTS
SELECT SUM(UIC_TOTAL) AS TOTAL_RESERVISTS
FROM   Area_G18_ZIP
WHERE  Area_G18_ZIP.UIC = MovingUnit.UIC
      AND Area_G18_ZIP.ZIPCODE = ANY (SELECT AREA_ZIPCODE.ZIP
                                       FROM AREA_ZIPCODE)

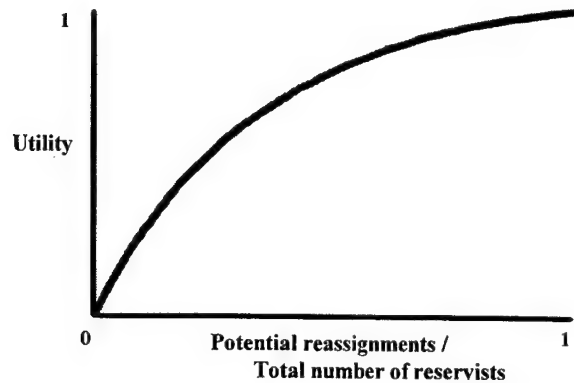
```

```

UIC_TOTAL
SELECT UIC_TOTAL
FROM   G18NatI
WHERE  G18NatI.UIC = MovingUnit.UIC

```

**Yield Curve:** The current location will always receive a utility score of 0.0 on this measure. For relatively close relocation sites, this function was made to be convex, assigning high utility values to alternatives that are close to the current location.



**Max Utility:** 1.0

**Min Utility:** 0.0

## Measure 16. Distance to Area Maintenance Support Activity

**Definition:** Distance to the nearest Area Maintenance Support Activity (AMSA) is calculated as a proxy measure for response time and support quality.

**Calculation:** The straight-line distance from the proposed site to the closest AMSA is calculated using a geocoded version of the AMSA file.

DIST\_AMSA[Determine distance to nearest AMSA Site]

**Units:** Miles

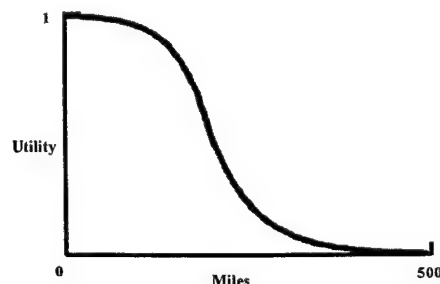
**Source Data:** AMSA

**ACROPOLIS Table(s):** NONE

**Query:** DIST\_AMSA(MapInfo)  
SELECT \*  
FROM AMSA  
WHERE Obj Withing ObjDistanceBuffer into TempRZA  
(Note: ObjDistanceBuffer is equal to 300 miles)

```
SELECT Distance((CentroidX(Obj), CentroidY(Obj), FacIDLat, FacIDLong, "mi")
FROM TempAMSA
ORDER BY Distance INTO TempAMSA.Dist
```

**Yield Curve:** The desirability of a relocation site is relatively insensitive to small changes in distance for both close and distant AMSA sites. Little degradation in service is expected if the AMSA can have parts and technicians on site within a couple hours using a car or truck. It is possible that a trainer that breaks down in the morning may be operational for an afternoon training session. At approximately 200 miles (assigned a 0.5 utility) it starts to become impractical to expect same day service and avoid an overnight stay. Eventually it becomes necessary to consider flying rather than driving which is likely to further reduce the responsiveness and effectiveness of the AMSA.



**Max Utility:** 0

**Min Utility:** > 500

THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 17. Distance to Nearest Equipment Concentration Site

**Definition:** Distance to the nearest Equipment Concentration Site (ECS) provides an indication of the training time that must be used to travel back and forth.

**Calculation:** The straight-line distance from the proposed site to the closest ECS is calculated using a geocoded version of the ECS file.

DIST\_ECS[Determine distance to nearest ECS]

**Units:** Miles

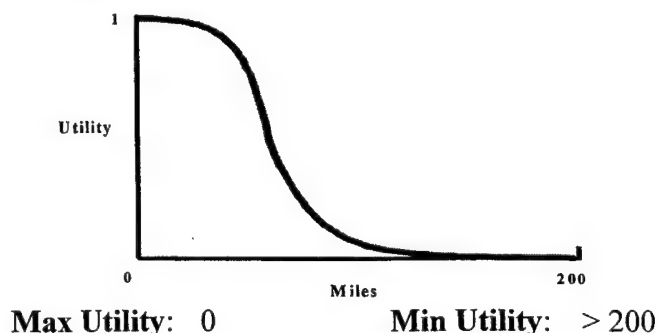
**Source Data:** ECS

**ACROPOLIS Table(s):** NONE

**Query:** DIST\_ECS(MapInfo)  
SELECT \*  
FROM ECS  
WHERE Obj Withing ObjDistanceBuffer into TempECS  
(Note: ObjDistanceBuffer is equal to 300 miles)

```
SELECT Distance((CentroidX(Obj), CentroidY(Obj), FacIDLat, FacIDLong, "mi")
FROM TempECS
ORDER BY Distance INTO TempECS.Dist
```

**Yield Curve:** The desirability of an Equipment Concentration Site is relatively insensitive to small changes in distance for both close and distant sites. Typically, a site that can be reached within an hour and ten minutes is not significantly less desirable than one that can be reached in ten minutes. An hour of one-way travel time is not normally considered to be excessive and allows for most of the time to be spent training on a one day training exercise. At approximately 60 miles (assigned a 0.5 utility) it starts to become impractical to expect useful training to be conducted on a day trip and avoid an overnight stay. Eventually it becomes necessary to consider flying rather than driving which is likely to further reduce the desirability of the ECS.



THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 18. Facility Weekends Used

**Definition:** Facility Weekend Usage provides the number of weekends per month that the facility is currently in use. This measure treats a facility as a limited resource that is incrementally depleted as more units are assigned. Since most units require exclusive use of the facility one weekend every month, the number of weekends used normally corresponds to the number of units assigned and is typically limited to four.

**Calculation:** This value is extracted from the "RS\_WKND\_PM" field of the COMPLEX file.

WKND\_USED[Retrieve Number Weekends Facility Used per Month]

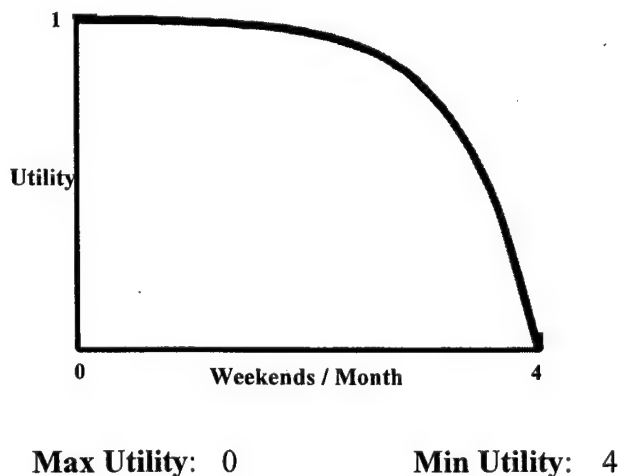
**Units:** Weekends per month

**Source Data:** COMPLEX

**ACROPOLIS Table(s):** COMPLEX\_

**Query:** WKND\_USED  
SELECT COMPLEX\_.FAC\_WKND\_USED  
FROM COMPLEX\_  
WHERE COMPLEX\_.FAC\_ID = ProposedFacility.FAC\_ID

**Yield Curve:** Although some exceptions exist, a typical facility offers no utility to a relocating unit if all four weekends are already being used. Although most facilities with three units or less should be able to accommodate a new unit and might be viewed as having equal utility, other issues such as full time administrative space and available equipment storage space make a facility with fewer units currently assigned slightly more desirable.





THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 19. Available MOS from Closing Units

**Definition:** This measure provides the number of reservists from closing units in the area of the proposed relocation site who possess a Military Occupational Specialty (MOS) needed by the relocating unit. These people provide a preferred pool of trained and qualified recruits.

**Calculation:** The number of personnel records (from the G18CWE file) that meet all the following requirements are counted:

1. The reservist is assigned to a unit that is scheduled to close (a TIER="5B" entry in the G17 file is used to produce a list of closing units).
2. The reservist lives in a zip code in the area of the proposed relocation site.
3. The reservist's primary MOS is needed by the moving unit.

If the three MOS groups with the largest number of members in the moving unit account for more than 50 percent of the total unit membership, then only those three MOS's are considered. Otherwise all MOS's required by the moving unit are considered as an MOS of interest.

TOTAL\_CLOS\_MOS[Total Number of Available Reservists from Area Closing Units with  
MOS's of Interest]

**Units:** Number of people

**Source File:** COMMAND PLAND, G17, G18CWE, GEOREF, US\_ZIPS(MapInfo)

**ACROPOLIS Table(s):** CMDPLAN, G17Natl, G18Natl, VALID\_UIC

**Query:** Area-FACID List(MapInfo)  
SELECT FAC\_ID INTO TempFACID  
FROM GEOREF  
WHERE Obj Within objAreaBuffer  
ORDER BY FAC\_ID  
(Note: objAreaBuffer is equal to 300 miles)

VALID\_UIC  
SELECT UIC, FAC\_ID, UnitName, City, State, Zip  
FROM G17Natl  
WHERE G17Natl.UIC = ANY (SELECT CMD\_PLAN.UIC  
FROM CMD\_PLAN)

Area-UIC List

```
SELECT DISTINCT UIC INTO AREA_UIC
FROM   VALID_UIC
WHERE  VALID_UIC.FAC_ID = ANY (SELECT AREA_FACID.FAC_ID
                                FROM AREA_FACID)
```

NoAssnxMOS

```
SELECT MOS, COUNT(*) AS MOS_COUNT INTO NoAssnxMOS
FROM   G18Natl
WHERE  G18Natl.UIC = MovingUnit.UIC
GROUP BY MOS
ORDER BY COUNT(*) DESC
```

MOS\_TOTAL

```
SELECT SUM(MOS_COUNT) AS MOS_TOTAL
FROM   NoAssnxMOS
```

MOS\_TOP3

```
SELECT TOP 3 MOS_COUNT
FROM   NoAssnxMOS
```

MOS\_INTEREST

```
IF MOS_TOP3/MOS_TOTAL < 50%
    SELECT MOS INTO MOS_INTEREST
    FROM   NoAssnxMOS
    ORDER BY MOS
IF MOS_TOP3/MOS_TOTAL > 50%
    SELECT TOP 3 MOS INTO MOS_INTEREST
    FROM   NoAssnxMOS
    ORDER BY MOS
```

AREA\_CLOS\_UIC

```
SELECT UIC
FROM   G17Natl
WHERE  G17Natl.TIER = "5B"
      AND G17Natl.UIC = ANY (SELECT AREA_UIC.UIC
                              FROM AREA_UIC)
```

AREA\_ZIPCODE(MapInfo)

```
SELECT ZIP_CODE AS ZIP
FROM   US_ZIPS
WHERE  Obj Within objAreaBuffer
ORDER BY ZIP_CODE
```

```

G18(MapInfo)
SELECT  UIC, LEFT$(ZIP,5) "ZIPCODE", PRI "MOS"
FROM    G18CWE
WHERE   Obj Within objG18Buffer AND PRI <> ""
ORDER BY  UIC, ZIP
INTO     G18

```

```

Area_G18_MOS
SELECT  DISTINCT UIC, ZipCode, MOS, COUNT(UIC) AS UIC_TOTAL
INTO    Area_G18_MOS
FROM    G18
GROUP BY  UIC, ZipCode, MOS
ORDER BY  UIC, ZipCode, MOS

```

```

Area_G18_ZIP
SELECT  DISTINCT UIC, ZIPCODE, COUNT(UIC) AS UIC_TOTAL
FROM    G18
GROUP BY  UIC, ZIPCODE
ORDER BY  UIC, ZIPCODE

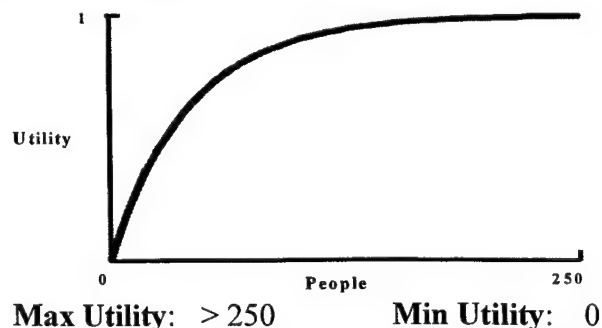
```

```

TOTAL_CLOS_MOS
SELECT  SUM(UIC_TOTAL) AS TOTAL_CLOS_MOS
FROM    Area_G18_MOS
WHERE   Area_G18_MOS.MOS = ANY (SELECT MOS_INTEREST.MOS
                                FROM MOS_INTEREST)
        AND Area_G18_ZIP.UIC = ANY (SELECT AREA_CLOS_UIC.UIC
                                FROM AREA_CLOS_UIC)
        AND Area_G18_ZIP.ZIPCODE = ANY (SELECT AREA_ZIPCODE.ZIP
                                FROM AREA_ZIPCODE)

```

**Yield Curve:** The shape of this function assumes diminishing returns on the number of transfers available. Experience suggests for an average unit of 100 people, that it is unusual to expect more than a third of the members to transfer from closing units with the proper MOS. Of the reservists in this category, only half typically transfer, so a value of 60 personnel is assigned a utility of 0.9. The incremental value added by each additional person over 60 continues to drop until no marginal gain is expected over 250.



THIS PAGE INTENTIONALLY LEFT BLANK

## Measure 20. Available MOS IRR

**Definition:** This measure provides the number of Individual Ready Reserve members who live in the area of the proposed relocation site and who possess a Military Occupational Specialty (MOS) needed by the relocating unit. These people provide a preferred pool of trained recruits.

**Calculation:** The number of IRR members who possess an MOS needed by the moving unit and who live in the area of the proposed relocation site (based upon the zip code of their home of record in the IRR file) are counted. If the three MOS groups with the largest number of members in the moving unit account for more than 50 percent of the total unit membership, then only those three MOSs are considered. Otherwise all MOSs required by the moving unit are considered as an MOS of interest.

TOTAL\_IRR\_MOS[Total Number of Available Reservists from the IRR with MOS's of Interest]

**Units:** Number of People

**Source File:** IRR, G18CWE

**ACROPOLIS Table(s):** G18NatI,

**Query:** NoAssnxMOS  
SELECT MOS, COUNT(\*) AS MOS\_COUNT INTO NoAssnxMOS  
FROM G18NatI  
WHERE G18NatI.UIC = MovingUnit.UIC  
GROUP BY MOS  
ORDER BY COUNT(\*) DESC

MOS\_TOTAL  
SELECT SUM(MOS\_COUNT) AS MOS\_TOTAL  
FROM NoAssnxMOS

MOS\_TOP3  
SELECT TOP 3 MOS\_COUNT  
FROM NoAssnxMOS

MOS\_INTEREST  
IF MOS\_TOP3/MOS\_TOTAL < 50%  
SELECT MOS INTO MOS\_INTEREST  
FROM NoAssnxMOS  
ORDER BY MOS

```

IF MOS_TOP3/MOS_TOTAL > 50%
  SELECT TOP 3 MOS INTO MOS_INTEREST
  FROM NoAssnxMOS
  ORDER BY MOS

```

```

IRR(MapInfo)
SELECT ZIPC "ZIP", LEFT$(PMOS, 3) "MOS"
FROM IRR
WHERE Obj Within objAreaBuffer and ZIPC <> "" AND PMOS <> ""
ORDER BY ZIPC
INTO IRR

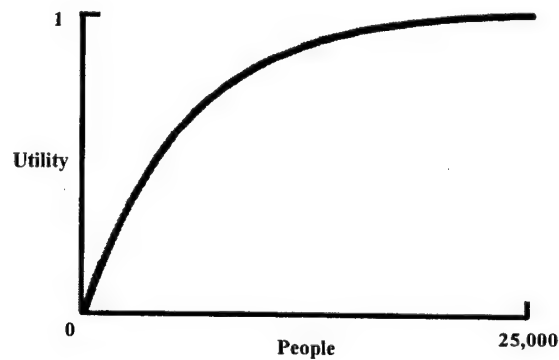
```

```

TOTAL_IRR_MOS
SELECT SUM(UIC_TOTAL) AS TOTAL_CLOS_MOS
FROM IRR
WHERE IRR.MOS = ANY (SELECT MOS_INTEREST.MOS
                      FROM MOS_INTEREST)

```

**Yield Curve:** IRR members represent preferred recruits for less than half of the positions of a typical moving unit (approximately 40 out of 100) because of issues such as seniority and changes in the skills associated with an MOS. The success rate of recruiting IRR members is approximately 1 out of 100, so 4000 IRR members in the area of the relocation site are required to provide sufficient market to fill the 40 positions. The value of 4000 is assigned the average utility value of 0.5. As the IRR market increases it exceeds the needs of the moving unit and makes the limited recruiting efforts marginally less effective.



**Max Utility:** > 25,000

**Min Utility:** 0

## APPENDIX B. ARIES SOURCE DATA FILE META-DATA

This appendix contains the meta-data that could be documented for the ARIES SDSS project source files. "ACROPOLIS" as used in this appendix refers to the file name of the ARIES data resource file.

### Index

1. AMSA.....	123
2. COMMAND PLAN.....	125
3. COMPLEX .....	127
4. ECS .....	129
5. FINANCE .....	131
6. FPS.....	133
7. FYxxLOSS.....	135
8. G17.....	137
9. G18CWE.....	139
10. G19TRUE .....	141
11. GEOREF .....	143
12. INTEREST.....	145
13. IRR.....	147
14. NGNON_CL .....	149
15. QMA .....	151
16. RPINFODT.....	153
17. RZA.....	155



THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: AMSA Location: ../Aries/MapBasic/USARCDData

File Type: FoxPro 2.6 Size(MB): .026 No. Records: 190

Associated ARIES Tables: Not in ACROPOLIS, Geocoded for use in MapInfo

### File Description:

AMSA File contains information about the location of each AMSA station. It is used in determining the value for the distance to the nearest AMSA.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
fac_id	Facility Identification Code	Char		No
fac_title	Facility Title	Char		No
fac_street	Street Address of Facility	Char		No
fac_city	City Facility is located in	Char		No
fac_state	State Facility is located in	Char		No
fac_zip	Zip Code of the Facility	Char		No
abb_type		Char		No

### Extract Queries:

NONE

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: COMMAND PLAN Location: ACROPOLIS  
 File Type: FoxPro 2.6 Size(MB): 3.29 No. Records: 9,897  
 Associated ARIES Tables: CMDPLAN

### File Description:

Command Plan is the file that contains information about each unit in the Army Reserve. It is used to cross reference FAC ID's with UIC's. It is also used to screen for Valid UIC's with in the next 13 months.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
UIC	Unit Identification Code	Char		yes
FACID	Facility Identification Code	Char		no
EDATE	Effective Date of Transaction	Char		no

### Extract Queries:

```

CMDPLAN
SELECT DISTINCT UIC, FACID AS FAC_ID,
    EDATE
FROM  COMMANDPLAN
WHERE (FACID <> "N/A") AND (FACID <>
    "TBD") AND (FACID <> "") AND
    (LEN(FACID) > 2) AND
    ((LEFT(EDATE,4) = '1998' AND
    MID(EDATE,5,2) <= '02') OR
    (LEFT(EDATE,4) <= '1997'))
ORDER BY  UIC, EDATE DESC
INTO  CMDPLAN
INDEX ON UIC as UIC
  
```

**Note:** Application automatically adjusts the dates to obtain a 13 month window.

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: COMPLEX Location: ACROPOLIS  
 File Type: FoxPro 2.6 Size(MB): 2.1 No. Records: 1,557  
 Associated ARIES Tables: COMPLEX\_

### File Description:

The Complex File is used to determine if the facility is owned by or leased to the government and the number of weekends each facility is used during a month.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
FAC_ID	Facility Identification Code	Char		yes
GOVT_OWN	Facility ownership status	Char	Y/N	no
RS_WKND_PM	Reserve Station weekend usage per mo.	Number	0-4	no

### Extract Queries:

```

COMPLEX_
SELECT FAC_ID, GOVT_OWN AS
      FAC_OWNED, RS_WKND_PM AS
      FAC_WKND_USED
FROM   COMPLEX
WHERE  LEN(FAC_ID) = 5
INTO   COMPLEX_
INDEX ON FAC_ID as FACID, Primary, Unique
  
```

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: ECS Location: ...\MapBasic\USARCDData\

File Type: FoxPro 2.6 Size(MB): .004 No. Records: 30

Associated ARIES Tables: Not in ACROPOLIS, Geocoded for use in MapInfo

### File Description:

ECS File contains information about the location of each Equipment Center. It is used in determining the distance to the nearest ECS.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
fac_id	Facility Identification Code	Char		No
fac_title	Facility Title	Char		No
fac_street	Street Address of Facility	Char		No
fac_city	City Facility is located in	Char		No
fac_state	State Facility is located in	Char		No
fac_zip	Zip Code of the Facility	Char		No
abb_type		Char		No

### Extract Queries:

NONE



THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: FINANCE Location: ACROPOLIS  
 File Type: FoxPro 2.6 Size(MB): 83.4 No. Records: 311,793  
 Associated ARIES Tables: FINANCE\_, FINANCE\_QTR

### File Description:

Finance is the file that contains pay information for the previous eight quarters about every Reservist. It is used to obtain information about Drill Attendance for a given Facility.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
CURR_UIC	Current Unit Identification Code	Char		No
UTA1QCFY	Unit Training Attendance for 1 <sup>st</sup> Qtr this FY	Number		No
UTA2QCFY	Unit Training Attendance for 2 <sup>nd</sup> Qtr this FY	Number		No
UTA3QCFY	Unit Training Attendance for 3 <sup>rd</sup> Qtr this FY	Number		No
UTA4QCFY	Unit Training Attendance for 4 <sup>th</sup> Qtr this FY	Number		No
UTA1Q1PF	Unit Training Attendance for 1 <sup>st</sup> Qtr last FY	Number		No
UTA2Q1PF	Unit Training Attendance for 2 <sup>nd</sup> Qtr last FY	Number		No
UTA3Q1PF	Unit Training Attendance for 3 <sup>rd</sup> Qtr last FY	Number		No
UTA4Q1PF	Unit Training Attendance for 4 <sup>th</sup> Qtr last FY	Number		No

### Extract Queries:

```

FINANCE_
SELECT "W" & LEFT(CURR_UIC,5) AS UIC,
      COUNT(CURR_UIC) AS
      UIC_TOTAL
FROM  FINANCE
WHERE CURR_UIC <> ""
ORDER BY CURR_UIC
GROUP BY CURR_UIC
INTO  FINANCE_
INDEX ON UIC as UIC
  
```

```

FINANCE_QTR
SELECT "W" & LEFT(CURR_UIC,5) AS UIC,
      UTA1QCFY, UTA2QCFY, UTA3QCFY,
      UTA4QCFY, UTA1Q1PF, UTA2Q1PF,
      UTA3Q1PF, UTA4Q1PF
FROM  FINANCE
WHERE CURR_UIC <> "" AND NPS_IND = NULL
      AND PAY_STAT = 'A'
ORDER BY CURR_UIC
INTO  FINANCE_QTR
INDEX ON UIC as UIC
  
```

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: FPS Location: ACROPOLIS

File Type: FoxPro 2.6 Size(MB): .088 No. Records: 1,561

Associated ARIES Tables: FPS

### File Description:

FPS is used to obtain information about the Cost to operate each facility as well as the Condition of each Facility. Used to return a value for the Cost per Square Foot and the Facility Condition.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
FAC_ID	Facility Identification Code	Char		No
FAC_COND	Condition of the Facility	Char		No
COST_PR_SF	Cost per Square Foot to Operate Facility	Number		No

### Extract Queries:

```
FPS_
SELECT FAC_ID, FAC_COND, COST_PR_SF
FROM   FPS
WHERE  FAC_ID <> ""
ORDER BY FAC_ID
INTO   FPS_
INDEX ON FAC_ID as FACID, Primary, Unique
```

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: FYxxLOSS Location: ACROPOLIS  
 File Type: FoxPro 2.6 Size(MB): 85.4 No. Records: 260,000  
 Associated ARIES Tables: FYxxLOSS, FYxxXFER

### File Description:

FYxxLOSS file contains information about the personnel losses incurred by each unit during a fiscal year.  
 It is used to determine the Average Loss and Transfer Rate of a Unit.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
UIC	Unit Identification Code	Char		No
TRMN	Transfer Reason Code	Char		No

### Extract Queries:

```

FYxxLOSS
SELECT UIC1 AS UIC, COUNT(UIC1) AS
      UIC_TOTAL
FROM   FY_LOSS
WHERE  TRMN = 'LOSS'
ORDER BY UIC1
GROUP BY UIC1
INTO   FYxxLOSS
INDEX ON UIC as UIC, Primary, Unique
  
```

```

FYxxXFER
SELECT UIC1 AS UIC, COUNT(UIC1) AS
      UIC_TOTAL
FROM   FY_LOSS
WHERE  TRMN = 'TRFD'
ORDER BY UIC1
GROUP BY UIC1
INTO   FYxxXFER
INDEX ON UIC as UIC, Primary, Unique
  
```

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name:     G17     Location:     ACROPOLIS      
 File Type:     FoxPro 2.6     Size(MB):     3.11     No. Records:     5,869      
 Associated ARIES Tables:     G17NatI    

### File Description:

G17 file contains facility Unitname, street address data and Zip Code. It is used as the primary cross reference with Command Plan to display facility information and validate user input.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
UIC	Unit Identification Code	Char		No
UNITNAME	Name of the Unit	Char		No
TCCCITY	City Unit is located in	Char		No
TCCSTATE	State Unit is located in	Char		No
TCCZIP	Zip code of the Unit	Char		No
TIER	Code used to determine if Unit is closing	Char		No
RECSTAT	Recruiting Station Code	Number		No
TYPEORG	Type of organization	Number		No

### Extract Queries:

```

G17NatI
SELECT UIC, UNITNAME, TCCCITY AS
      CITY, TCCSTAT AS STATE,
      LEFT(TCCZIP,5) AS ZIP, TIER
FROM   G17
WHERE  (RECSTAT < "1") AND (TYPEORG
      < "2") AND UIC < ""
ORDER BY UIC
INTO   G17NatI
INDEX ON UIC as UIC, Primary, Unique
  
```



THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: G18CWE Location: ACROPOLIS;..\MapBasic\UsarcData  
 File Type: FoxPro 2.6 Size(MB): 145.9 No. Records: 208,416  
 Associated ARIES Tables: G18NatI, G18NatI\_UIC; also Geocoded for use in MapInfo

### File Description:

G18 File contains information about personnel in the US Army Reserves. It is used in determining the Total Number Assigned used in calculating the Loss/Transfer Rates, Total Available Closing and the Reassignments values. Also used to obtain a list of the Zip Code's and MOSs of every Reservists with their associated UIC..

### Required Data Elements

Name	Description	Data Type	Format	Key Field
UIC	Unit Identification Code assigned	Char		No
ZIP	Zip Code of the individual	Char		No
PRI	Primary MOS	Char		No

### Extract Queries:

```
G18NatI
SELECT UIC, LEFT(ZIP,5) AS ZIPCODE,
      LEFT(PRI,3) AS MOS
FROM G18_
WHERE PRI <> "" AND UIC <> ""
ORDER BY UIC
INTO G18NatI
INDEX ON UIC as UIC
```

```
G18NatI_UIC
SELECT UIC, COUNT(UIC) AS UIC_TOTAL
FROM G18NatI
ORDER BY UIC
GROUP BY UIC
INTO G18NatI_UIC
INDEX ON UIC as UIC, Primary, Unique
```

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: G19TRUE Location: ACROPOLIS  
 File Type: FoxPro 2.6 Size(MB): 14.4 No. Records: 233,211  
 Associated ARIES Tables: G19NatI

### File Description:

G19 File contains information about the required manning levels of each Unit. It is used in determining Average Area Manning for a Facility.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
OWN_UIC	Unit Identification Code	Char		No

### Extract Queries:

```

G19NatI
SELECT OWN_UIC AS UIC,
       COUNT(OWN_UIC) AS
       UIC_TOTAL
FROM   G19
WHERE  OWN_UIC <> ""
ORDER BY OWN_UIC
GROUP BY OWN_UIC
INTO   G19NatI
INDEX ON UIC as UIC
    
```

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name:       GEOREF       Location:       ACROPOLIS;..\MapBasic\UsarcData      

File Type:       FoxPro 2.6       Size(MB):       .21       No. Records:       1,553      

Associated ARIES Tables:       VALID\_UNIT; also Geocoded for use in MapInfo      

### File Description:

Georef File contains specific information about each Unit. It is used to verify and cross reference FACID's and UIC as well as Facility and Unit specific information.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
FAC_ID	Facility Identification Code	Char		No
FAC_TITLE	Name of the Facility	Char		No
FAC_CITY	City the Facility is located in	Char		No
FAC_STATE	State the Facility is located in	Char		No
FAC_ZIP	Zip Code of the Facility	Char		No
Latitude	Position of Facility by degree of latitude	Number		No
Longitude	Position of Facility by degree of longitude	Number		No

### Extract Queries:

```
VALID_UNIT
SELECT FAC_ID, FAC_TITLE AS
      UNITNAME, FAC_CITY AS CITY,
      FAC_STATE AS STATE,
      LEFT(FAC_ZIP,5) AS ZIP
FROM   GEOREF
WHERE  FAC_ID <> ""
ORDER BY FAC_ID
INTO   VALID_UNIT
INDEX ON FAC_ID as FACID
```

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: INTEREST Location: ACROPOLIS  
 File Type: FoxPro 2.6 Size(MB): 4.2 No. Records: 3,985  
 Associated ARIES Tables: INTEREST

### File Description:

Interest File contains information about facilities and the date they were acquired. It is used to calculate the Facility Age for each facility..

### Required Data Elements

Name	Description	Data Type	Format	Key Field
FAC_IDSTR	Facility Identification Code	Char		No
DATE_ACQ	Date Facility Acquired	Date		No
ABB_TYPE		Char		No

### Extract Queries:

```

INTEREST_
SELECT FAC_IDSTR AS FAC_ID, DATE_ACQ
FROM INTEREST
WHERE FAC_IDSTR <> "" AND ABB_TYPE =
      "USARC (MB)" AND NOT
      ISNULL(DATE_ACQ)
ORDER BY FAC_IDSTR
INTO INTEREST_
INDEX ON FAC_ID as FACID, Primary, Unique
    
```



THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: IRR Location: ...\MapBasic\UsarcData

File Type: FoxPro 2.6 Size(MB): 7.5 No. Records: 140,077

Associated ARIES Tables: Not in ACROPOLIS, Geocoded for use in MapInfo

### File Description:

IRR File contains information about the individuals listed in the Individual Ready Reserve. It is used to determine the value for IRR Available and Available MOS IRR.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
ZIPC	Zip Code for IRR Individual	Char		No

### Extract Queries:

NONE

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: NGNON\_CL Location: ...\MapBasic\UsarcData

File Type: FoxPro 2.6 Size(MB): .64 No. Records: 3,673

Associated ARIES Tables: Not in ACROPOLIS, Geocoded for use in MapInfo

### File Description:

NGNON\_CL File contains information about the non-closing National Guard Units. It is used in determining the value for Competition.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
UPC		Char		Yes
ZIP	Zip Code for National Guard Individual	Char		Yes

### Extract Queries:

NONE

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: QMA Location: ...\MapBasic\UsarcData  
 File Type: FoxPro 2.6 Size(MB): 2.8 No. Records: 34,265  
 Associated ARIES Tables: Not in ACROPOLIS, Geocoded for use in MapInfo

### File Description:

QMA File contains Census information. It is used in determining the value for Recruit Market for each Facility.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
ZIP	Zip Code	Char		No
MWCAT12	White Male Mental Categories 1 &2	Number		No
MWCAT3A	White Male Mental Category 3A	Number		No
MBCAT12	Black Male Mental Categories 1 &2	Number		No
MBCAT3A	Black Male Mental Category 3A	Number		No
MHCAT12	Hispanic Male Mental Categories 1 &2	Number		No
MHCAT3A	Hispanic Male Mental Category 3A	Number		No

### Extract Queries:

NONE

THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name: RPINFODT Location: ACROPOLIS  
 File Type: FoxPro 2.6 Size(MB): 14.3 No. Records: 47,159  
 Associated ARIES Tables: FPINFODT\_

### File Description:

RPINFODT is a file that contains information about the backlogged maintenance costs of each Facility. It is used to determine the amount of backlogged maintenance is required at the given Facility.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
FAC_ID	Facility Identification	Char		No
CWE_TOTAL	Total amount of outstanding Maint. Actions	Number		No

### Extract Queries:

```

RPINFODT_
SELECT FAC_ID, SUM(CWE_TOTAL) AS
    MAINT_COST
FROM   RPINFODT
WHERE  FAC_ID <> ""
ORDER BY FAC_ID
GROUP BY FAC_ID
INTO   RPINFODT_
INDEX ON FAC_ID as FACID, Primary, Unique
    
```



THIS PAGE LEFT INTENTIONALLY BLANK

## A R I E S Data File Documentation Form

ARIES File Name:     RZA     Location:     ...\MapBasic\UsarcData    

File Type:     FoxPro 2.6     Size(MB):     .16     No. Records:     1,793    

Associated ARIES Tables:     Not in ACROPOLIS, Geocoded for use in MapInfo    

### File Description:

RZA File contains information about the location of Recruit Stations. It is used to determine the distance to the nearest Recruit Station.

### Required Data Elements

Name	Description	Data Type	Format	Key Field
rsid	Recruit Station Identification Code	Char		No
name	Recruit Station Title	Char		No
zip	Zip Code of the Recruit Station	Char		No
latitude	Position of Recruit Station by latitude	Number		No
longitude	Position of Recruit Station by longitude	Number		No

### Extract Queries:

NONE

THIS PAGE LEFT INTENTIONALLY BLANK

## APPENDIX C. ARIES DECISION MEASURE STATISTICS

This Appendix contains the statistics calculated for the USARC data set. Validity and Frequency Statistics were calculated for 17 of the 20 measures that were not dependent on knowing the identification of the Moving Unit.

Individual percentages in the frequency distributions for the 17 measures in this appendix are percentages relative to total non-missing values.

The limits of the ranges for valid values were determined by using a rule of reasonableness to identify values that would adversely affect the evaluation process. Consideration was given to the following areas; (1) the range of values returned during the evaluation process, (2) expected values based on the Yield Curves and (3) common sense (i.e., 0 value).

### Index

ARIES Descriptive Statistics .....	159
ARIES Measures Analysis.....	161
Measure 1. Facility Backlogged Maintenance.....	163
Measure 2. Facility Operating Costs.....	165
Measure 3. Facility Age.....	167
Measure 4. Facility Condition.....	169
Measure 5. Facility Ownership.....	171
Measure 6. Competition.....	173
Measure 7. Average Area Drill Attendance.....	175
Measure 8. Area Loss Rate .....	177
Measure 9. Area Transfer Rate .....	179
Measure 10. Area Average Manning .....	181
Measure 11. Distance to Nearest Recruit Station .....	183
Measure 12. Available Transfers from Closing Units .....	185
Measure 13. IRR Available.....	187
Measure 14. Recruit Market.....	189
Measure 16. Distance to Nearest Area Maintenance Support Activity .....	191
Measure 17. Distance to Nearest Equipment Concentration Site .....	193
Measure 18. Facility Weekends Used.....	195
Query Time .....	197

THIS PAGE LEFT INTENTIONALLY BLANK

**ARIES Descriptive Statistics**  
1325 U.S. Army Reserve Facilities

Measure	Observations (N)	Min Value	Max Value	Mean	Std Deviation
1 Facility Backlogged Maint.	1,205	0	11,979,371	448,131	837,391
2 Facility Operating Cost	1,251	0.0	293.5	3.0865	9.7124
3 Facility Age	765	0	1,677	295.1	173.3
4 Facility Condition	1,251	N/A	N/A	N/A	N/A
5 Facility Owned	1,319	N/A	N/A	N/A	N/A
6 Competition	1,300	18	20,759	4,116.3	3,960.0
7 Area Drill Attendance	1,300	0.20	0.82	0.58	0.06
8 Area Loss Rate	1,325	0.00	0.86	0.32	0.11
9 Area Transfer Rate	1,300	0.00	1.84	0.27	0.20
10 Area Average Manning	1,325	0.00	1.94	0.86	0.20
11 Distance to Recruiter	1,325	0	7,619.9	18.2	287.7
12 Area Avail Closing Unit	819	1	504	75.2	114.1
13 IRR Available	1,315	1	3,497	395.8	658.9
14 Area Recruit Market	1,316	253	214,738	33,189.9	41,290.4
15 *Reassignments					
16 Distance to AMSA	1,325	0	7,619.9	42.4	289.1
17 Distance to ECS	1,325	0	5,290.9	268.1	510.1
18 Facility Weekends Used	1,320	0	3	1.6	1.0
19 *Avail MOS Closing Units					
20 *Available MOS IRR					

\* Moving Unit Specific Measures

(Minutes)	Observations (N)	Min	Max	Mean	Std Deviation
Time to Complete Queries	1325	1.7	76.1	8.7	6.7

THIS PAGE LEFT INTENTIONALLY BLANK

**ARIES Measures Analysis**  
1325 U.S. Army Reserve Facilities

Measure	Missing	Out of Range	Potentially Valid	Valid Range
1 Facility Backlogged Maint.	9.1%	1.7%	89.2%	$0 < x_1 < 20M$
2 Facility Operating Cost	5.6%	18.4%	76.0%	$0 < x_2 < 100$
3 Facility Age	42.3%	0.2%	57.6%	$x_3 > 0$
4 Facility Condition	5.6%	0.0%	94.4%	$x_4 = G \text{ or } A \text{ or } R$
5 Facility Owned	0.5%	0.0%	99.5%	$x_5 = Y \text{ or } N$
6 Competition	1.9%	0.0%	98.1%	$0 < x_6 < 21,000$
7 Area Drill Attendance	1.9%	0.0%	98.1%	$0 \leq x_7 < 1$
8 Area Loss Rate	0.0%	2.1%	97.9%	$0 < x_8 < 1.0$
9 Area Transfer Rate	1.9%	1.1%	97.0%	$0 < x_9 < 1.0$
10 Area Average Manning	0.0%	2.3%	97.7%	$0 < x_{10} < 1.5$
11 Distance to Recruiter	0.0%	0.2%	99.8%	$x_{11} > 500$
12 Area Available Closing Unit	38.2%	0.0%	61.8%	$x_{12} \geq 0$
13 IRR Available	0.8%	0.0%	99.2%	$x_{13} > 0$
14 Area Recruit Market	0.7%	0.0%	99.3%	$x_{14} > 0$
15 *Reassignments				
16 Distance to AMSA	0.0%	0.2%	99.8%	$x_{16} > 500$
17 Distance to ECS	0.0%	10.6%	89.4%	$x_{17} > 500$
18 Facility Weekends Used	0.4%	0.0%	99.6%	$x_{18} > 4$
19 *Available MOS Closing Units				
20 *Available MOS IRR				

\* Moving Unit Specific Measures



THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 1. Facility Backlogged Maintenance

Frequency Data				
Values (Millions)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	23	1.9	23	1.9
> 0 - .5	850	70.5	873	72.4
> .5 - 1	198	16.4	1071	88.9
> 1 - 1.5	80	6.6	1151	95.5
> 1.5 - 2	27	2.2	1178	97.8
> 2 - 3	17	1.4	1195	99.2
> 3 - 4	2	0.2	1197	99.3
> 4 - 5	0	0.0	1197	99.3
> 5 - 10	5	0.4	1202	99.8
> 10 - 15	3	0.2	1205	100.0
> 15 - 20	0	0.0	1205	100.0
> 20	0	0.0	1205	100.0
<b>Total Non-Missing</b>	1205			
<b>Missing</b>	120			
<b>Total</b>	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1205	0	11,979,371	448,130.8	837,390.9

**Max Utility:** 0

**Min Utility:** 1,000,000

**Databases:** RPINFODT

THIS PAGE LEFT INTENTIONALLY BLANK

## Measure 2. Facility Operating Costs

Frequency Data				
Values (Millions)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	242	19.3	242	19.3
> 0 - 2	465	37.2	707	56.5
> 2 - 4	305	24.4	1012	80.9
> 4 - 6	97	7.8	1109	88.6
> 6 - 8	39	3.1	1148	91.8
> 8 - 10	44	3.5	1192	95.3
> 10 - 20	41	3.3	1233	98.6
> 20 - 50	16	1.3	1249	99.8
> 50 -100	0	0.0	1249	99.8
> 100 - 200	1	0.1	1250	99.9
> 200	1	0.1	1251	100.0
<b>Total Non-Missing</b>	1251			
<b>Missing</b>	74			
<b>Total</b>	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1251	0	294	3.1	9.7

**Max Utility:** 0

**Min Utility:** 100

**Databases:** FPS

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 3. Facility Age

Frequency Data				
Values (Months)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
< 0	0	0.0	0	0.0
0	2	0.3	2	0.3
0 - 100	81	10.6	83	10.8
101 - 200	33	4.3	116	15.2
201 - 300	361	47.2	477	62.4
301 - 400	29	3.8	506	66.1
401 - 500	185	24.2	691	90.3
501 - 750	70	9.2	761	99.5
751 - 1000	1	0.1	762	99.6
1001 - 1500	2	0.3	764	99.9
1501 - 2000	1	0.1	765	100.0
> 2000	0	0.0	765	100.0
Total Non-Missing	765			
Missing	560			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
765	0	1,677	295.1	173.3

Max Utility: 0

Min Utility: 1,200

Databases: INTEREST

THIS PAGE LEFT INTENTIONALLY BLANK

#### Measure 4. Facility Condition

Frequency Data				
Values	Frequency	Percent	Cumulative Frequency	Cumulative Percent
GREEN	1251	100.0	1251	106.3
AMBER	0	0.0	1251	106.3
RED	0	0.0	1251	106.3
Total Non-Missing	1251			
Missing	74			
Total	1251			

Green Utility: 1.0

Amber Utility: 0.5

Red Utility: 0.0

Databases: FPS



THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 5. Facility Ownership

Frequency Data				
Values	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Y	1110	83.8	1110	84.2
N	209	15.8	1319	100.0
Total Non-Missing	1319			
Missing	6			
Total	1325			

**Max Utility:** YES

**Min Utility:** NO

**Databases:** COMPLEX

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 6. Competition

Frequency Data				
Values (People)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	0	0.0	0	0.0
0 - 1,000	309	23.8	309	23.8
1,001 - 2,000	210	16.2	519	39.9
2,001 - 3,000	166	12.8	685	52.7
3,001 - 4,000	129	9.9	814	62.6
4,001 - 5,000	52	4.0	866	66.6
5,001 - 7,500	221	17.0	1087	83.6
7,501 - 10,000	80	6.2	1167	89.8
10,001 -15,000	99	7.6	1266	97.4
10,001 - 20,000	31	2.4	1297	99.8
> 20,000	3	0.2	1300	100.0
Total Non-Missing	1300			
Missing	25			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1300	18	20,759	4,116.3	3,960.0

Max Utility: 0

Min Utility: 10,000

Databases: COMMAND PLAN, G17, G19TRUE, GEOREF, NGNON\_CL

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 7. Average Area Drill Attendance

Frequency Data				
Values (Percent)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	0	0.0	0	0.0
0.01 - 0.10	0	0.0	0	0.0
0.11 - 0.20	0	0.0	0	0.0
0.21 - 0.30	4	0.3	4	0.3
0.31 - 0.40	14	1.1	18	1.4
0.41 - 0.50	74	5.7	92	7.1
0.51 - 0.60	703	54.1	795	61.2
0.61 - 0.70	472	36.3	1267	97.5
0.71 - 0.80	32	2.5	1299	99.9
0.81 - 0.90	1	0.1	1300	100.0
0.91 - 0.99	0	0.0	1300	100.0
>= 1.0	0	0.0	1300	100.0
<b>Total Non-Missing</b>	1300			
<b>Missing</b>	25			
<b>Total</b>	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1300	0.20	0.82	0.58	0.06

Max Utility: 1

Min Utility: 0

Databases: COMMAND PLAN, FINANCE, G17, G19TRUE, GEOREF

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 8. Area Loss Rate

Frequency Data				
Values (Percent)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	28	2.1	28	2.1
0.01 - 0.10	11	0.8	39	2.9
0.11 - 0.20	44	3.3	83	6.3
0.21 - 0.30	494	37.3	577	43.5
0.31 - 0.40	555	41.9	1132	85.4
0.41 - 0.50	144	10.9	1276	96.3
0.51 - 0.60	21	1.6	1297	97.9
0.61 - 0.70	17	1.3	1314	99.2
0.71 - 0.80	7	0.5	1321	99.7
0.81 - 0.90	4	0.3	1325	100.0
0.91 - 0.99	0	0.0	1325	100.0
>= 1.0	0	0.0	1325	100.0
<b>Total Non-Missing</b>	1325			
<b>Missing</b>	0			
<b>Total</b>	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1325	0.00	0.86	0.32	0.11

Max Utility: 0

Min Utility: 1

Databases: COMMAND PLAN, FYxxLOSS, G17, G18CWE, GEOREF



THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 9. Area Transfer Rate

Frequency Data				
Values (Percent)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	7	0.5	7	0.5
0.01 - 0.10	193	14.8	200	15.4
0.11 - 0.20	379	29.2	579	44.5
0.21 - 0.30	279	21.5	858	66.0
0.31 - 0.40	224	17.2	1082	83.2
0.41 - 0.50	71	5.5	1153	88.7
0.51 - 0.60	44	3.4	1197	92.1
0.61 - 0.70	34	2.6	1231	94.7
0.71 - 0.80	15	1.2	1246	95.8
0.81 - 0.90	42	3.2	1288	99.1
0.91 - 0.99	4	0.3	1292	99.4
>= 1.0	8	0.6	1300	100.0
Total Non-Missing	1300			
Missing	25			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1300	0.00	1.84	0.27	0.20

Max Utility: 0

Min Utility: >= 1

Databases: COMMAND PLAN, FYxxLOSS, G17, G18CWE, GEOREF

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 10. Area Average Manning

Frequency Data				
Values (Percent)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	25	1.9	25	1.9
0.01 - 0.50	22	1.7	47	3.5
0.51 - 0.75	183	13.8	230	17.4
0.76 - 0.80	124	9.4	354	26.7
0.81 - 0.90	355	26.8	709	53.5
0.91 - 1.00	455	34.3	1164	87.8
1.01 - 1.20	137	10.3	1301	98.2
1.21 - 1.40	18	1.4	1319	99.5
1.41 - 1.60	0	0.0	1319	99.5
1.61 - 1.80	5	0.4	1324	99.9
1.81 - 1.90	0	0.0	1324	99.9
1.91 - 2.0	1	0.1	1325	100.0
> 2.0	0	0.0	1325	100.0
Total Non-Missing	1325			
Missing	0			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1325	0.00	1.94	0.86	0.20

Max Utility: 1.25

Min Utility: 0

Databases: COMMAND PLAN,G17, G18CWE, G19TRUE, GEOREF

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 11. Distance to Nearest Recruit Station

Frequency Data				
Values (Miles)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2	0.2	2	0.2
> 0 - 10	1098	82.9	1100	83.0
> 10 - 20	110	8.3	1210	91.3
> 20 - 30	76	5.7	1286	97.1
> 30 - 40	25	1.9	1311	98.9
> 40 - 50	7	0.5	1318	99.5
> 50 - 75	3	0.2	1321	99.7
> 75 - 100	2	0.2	1323	99.8
> 100 - 200	0	0.0	1323	99.8
> 200 - 300	0	0.0	1323	99.8
> 300	2	0.2	1325	100.0
Total Non-Missing	1325			
Missing	0			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1325	0.0	7,619.9	18.2	287.7

Max Utility: 0

Min Utility: >= 100

Databases: RZA

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 12. Available Transfers from Closing Units

Frequency Data				
Value (People)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	0	0.0	0	0.0
1 - 50	509	62.1	509	62.1
51 - 100	143	17.5	652	79.6
101 - 150	44	5.4	696	85.0
151 - 200	40	4.9	736	89.9
201 - 250	7	0.9	743	90.7
251 - 300	12	1.5	755	92.2
301 - 350	15	1.8	770	94.0
351 - 400	7	0.9	777	94.9
401 - 500	41	5.0	818	99.9
> 500	1	0.1	819	100.0
Total Non-Missing	819			
Missing	506			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
819	1	504	75.2	114.1

Max Utility: 250

Min Utility: 0

Databases: COMMAND PLAN, G17, G18CWE, GEOREF



THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 13. IRR Available

Frequency Data				
Values (People)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	0	0.0	0	0.0
0 - 50	425	32.3	425	32.3
51 - 100	171	13.0	596	45.3
101 - 150	101	7.7	697	53.0
151 - 200	123	9.4	820	62.4
201 - 250	39	3.0	859	65.3
251 - 500	172	13.1	1031	78.4
501 - 1,000	132	10.0	1163	88.4
1,001 - 2,000	86	6.5	1249	95.0
2,001 - 3,000	44	3.3	1293	98.3
> 3,000	22	1.7	1315	100.0
Total Non-Missing	1315			
Missing	10			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1315	1	3,497	395.8	658.9

Max Utility:  $\geq 10,000$

Min Utility: 0

Databases: IRR

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 14. Recruit Market

Frequency Data				
Values (People)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	0	0.0	0	0.0
0 - 5,000	249	18.9	249	18.9
5,001 - 10,000	230	17.5	479	36.4
10,001 - 25,000	296	22.5	775	58.9
25,001 - 50,000	281	21.4	1056	80.2
50,001 - 100,000	159	12.1	1215	92.3
100,001 - 150,000	57	4.3	1272	96.7
150,001 - 200,000	26	2.0	1298	98.6
200,001 - 250,000	18	1.4	1316	100.0
250,001 - 300,000	0	0.0	1316	100.0
> 300,000	0	0.0	1316	100.0
Total Non-Missing	1316			
Missing	9			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1316	253	214,738	33,189.9	41,290.4

Max Utility:  $\geq 250,000$

Min Utility: 0

Databases: QMA

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 16. Distance to Nearest Area Maintenance Support Activity

Frequency Data				
Values (Miles)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	9	0.7	9	0.7
> 0 - 10	507	38.3	516	38.9
> 10 - 20	178	13.4	694	52.4
> 20 - 30	115	8.7	809	61.1
> 30 - 40	122	9.2	931	70.3
> 40 - 50	83	6.3	1014	76.5
> 50 - 75	147	11.1	1161	87.6
> 75 - 100	107	8.1	1268	95.7
> 100 - 200	48	3.6	1316	99.3
> 200 - 300	5	0.4	1321	99.7
> 300	4	0.3	1325	100.0
Total Non-Missing	1325			
Missing	0			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1325	0.0	7,619.9	42.4	289.1

Max Utility: 0

Min Utility: >= 500

Databases: AMSA

THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 17. Distance to Nearest Equipment Concentration Site

Frequency Data				
Values (Miles)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	84	6.3	84	6.3
> 0 - 10	43	3.2	127	9.6
> 10 - 20	38	2.9	165	12.5
> 20 - 30	45	3.4	210	15.8
> 30 - 40	54	4.1	264	19.9
> 40 - 50	41	3.1	305	23.0
> 50 - 75	129	9.7	434	32.8
> 75 - 100	151	11.4	585	44.2
> 100 - 200	401	30.3	986	74.4
> 200 - 300	178	13.4	1164	87.8
> 300	161	12.2	1325	100.0
<b>Total Non-Missing</b>	1325			
<b>Missing</b>	0			
<b>Total</b>	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1325	0.0	5,290.9	268.1	510.1

**Max Utility:** 0

**Min Utility:** >= 200

**Databases:** ECS



THIS PAGE LEFT INTENTIONALLY BLANK

### Measure 18. Facility Weekends Used

Frequency Data				
Values	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	237	18.0	237	18.0
1	430	32.6	667	50.5
2	331	25.1	998	75.6
3	322	24.4	1320	100.0
4	0	0.0	1320	100.0
Total Non-Missing	1320			
Missing	5			
Total	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1320	0	3	1.6	1.0

Max Utility: 0

Min Utility: 4

Databases: COMPLEX

THIS PAGE LEFT INTENTIONALLY BLANK

### Query Time

Frequency Data				
Values (Minutes)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
> 0 - 2	7	0.5	7	0.5
> 2 - 4	193	14.6	200	15.1
> 4 - 6	343	25.9	543	41.0
> 6 - 8	239	18.0	782	59.0
> 8 - 10	175	13.2	957	72.2
> 10 - 12	129	9.7	1086	82.0
> 12 - 15	117	8.8	1203	90.8
> 15 - 20	50	3.8	1253	94.6
> 20 - 30	44	3.3	1297	97.9
> 30	28	2.1	1325	100.0
<b>Total</b>	1325			

Descriptive Statistics				
N	Min	Max	Mean	Std Dev
1325	1.7	76.1	8.7	6.7

THIS PAGE LEFT INTENTIONALLY BLANK

## **APPENDIX D. SOURCE FILE DOCUMENTATION FORMS**

This appendix contains the recommended forms to be used in gathering meta-data information for data files used in conjunction with developing a SDSS application.

THIS PAGE LEFT INTENTIONALLY BLANK

## Source File Documentation Form

Customer Path & File Name: _____		LAN Server: _____	
Point of Contact(Office & Phone): _____			
File Type: _____		Size(MB): _____	No. Records: _____
Source File Name: _____		Update Frequency: _____	
Source File Location: _____	POC: _____	Phone: _____	
Application File Name: _____			
<b>File Description:</b>          			
<b>Queries Used:</b>          			



## Source File Documentation Form

Source File Name: \_\_\_\_\_ Location: \_\_\_\_\_

File Type: \_\_\_\_\_ Size(MB): \_\_\_\_\_ No. Records: \_\_\_\_\_

Associated Tables: \_\_\_\_\_

### File Description:

### Required Data Elements

Name	Description	Data Type	Format	Key Field

### Extract Queries:

## LIST OF REFERENCES

1. Murphy, Mark A., *An Automated Spatial Decision Support System for the Relocation of Army Reserve Units*, Master's Thesis, Naval Postgraduate School, March 1997.
2. Thomas, G. and Dolk, D. and Murphy, M., *ARIES: Army Reserve Installation Evaluation System*, Technical Report, Naval Postgraduate School, May 1997.
3. Brackett, Michael H., *The Data Warehouse Challenge*, John Wiley & Sons, Inc., New York, NY, 1996.
4. MapInfo Consulting Services, *MapInfo and the Data Warehouse*, MapInfo Corporation, 1996.
5. Environmental Systems Research Institute, Inc., *Spatial Data Warehouse*, Environmental Systems Research Institute, Inc., 1997.
6. Demarest, Marc, "Building the Data Mart", *DBMS Magazine*, Vol. 7, No. 8, URL: <http://visa.hevanet.com/demarest/marc/marts.html>, November 1993.
7. Inmon, William H. and Hackathorn, Richard D., *Using the Data Warehouse*, John Wiley & Sons, Inc., New York, NY, 1994.
8. McElreath, Jack, *Data Warehouses: An Architectural Perspective*, Computer Sciences Corporation, 1995.
9. Kimball, Ralph, "A Dimensional Modeling Manifesto", *DBMS Magazine*, Vol. 10, No. 9, URL: <http://www.dbmsmag.com/9708d15.html>, August 1997
10. Darling, Charles B., "Manage Your Reporting Environment", *Datamation Magazine*, May 1996.
11. Bersin, Josh, *Data Marts and Beyond: A Pragmatic Approach to Enterprise Decision Support*, International Data Warehousing Association and Onward Technologies, Inc., 1997.
12. Inmon, William H., *What is a Data Mart?*, D2K Incorporated, 1996.
13. Hufford, Duane, *Data Warehouse Quality: Part I*, Data Warehouse Resources, URL: <http://www.data-warehouse.com/resource/articles>, June 1997.
14. Orr, Ken, *Data Quality and Systems Theory*, The Ken Orr Institute, 1997.
15. Wand, Yang and Wang, Richard., "Anchoring Data Quality Dimensions inn Ontological Foundations", *Communications of the ACM*, Vol. 39, No. 11, November 1996.
16. Defense Information Systems Agency, *DoD Guidelines on Data Quality Management (Summary)*, URL: <http://164.117.192.240/srp/dqpaper.html>, August 1996.
17. Atkins, Mark E., "WARNING: Failure to know what's in legacy data may undermine your Data Warehouse", *Data Warehouse Institute Journal*, Vol. 2, 1997.

18. English, Larry P., "Help For Data Quality", *Information Week* (I 600), October 7, 1996.
19. Celko, Joe and McDonald, Jackie, "Don't Warehouse Dirty Data", *Datamation Magazine*, October 1995.
20. Bohn, Kathy, "Converting Data for Warehouses", *DBMS Online Magazine*, June 1997.

## BIBLIOGRAPHY

- Boar, Bernard, Understanding Data Warehousing Strategically, NCR Communications White Paper, URL: <http://www.tekptr.com/tpi/tdwi/review/bboar1.html>.
- Bruce, Thomas A., *Designing Quality Databases with IDEF1X Information Models*, Dorset House Publishing, New York, NY, 1992.
- Dolk, Daniel R. et al., Combining a Decision Model and GIS for Site Location Problems, Naval Postgraduate School White Paper, CodeSM/Dk, Monterey, Calif., 93943, October 1996.
- Fayyad, Usama M., "Data Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert Systems & their Applications*, Vol. 11, No. 5, October 1996.
- Foley, John, "Data Warehouse Pitfalls – Avoid common missteps by assuring data quality and focusing on business objectives", *Information Week*, Issue 631, May 19, 1997.
- Hansen, George W. et al., *Database Management and Design*, Prentice-Hall Inc., Englewood Cliffs, NJ, 1992.
- Kimball, Ralph, "Dealing with Dirty Data", *DBMS Online Magazine*, URL: <http://www.dbms.mti.com/9609d14.html>, September 1996.
- Leong-Hong, Belkis W. et al., *Data Dictionary/Directory Systems: Administration, Implementation and Usage*, John Wiley & Sons Inc., New York, NY, 1982.
- Mesrobian, Edmond, et al., "Mining Geophysical Data for Knowledge", *IEEE Expert Intelligent Systems & their Applications*, Vol. 11, No. 5, October 1996.
- Moore, J. H., et al., *Design of Decision Support Systems*, Data Base, Vol. 12, Nos. 1 and 2, Fall 1980.
- Orman, Levant, et al., Systems Approaches to Improving Data Quality, Massachusetts Institute of Technology White Paper TDQM-94-05, Cambridge, MA, URL: <http://web.mit.edu/tdqm/www/papers/94/95-05.html>, August 1994.
- Orr, Ken, Data Warehousing Technology, The Ken Orr Institute Whit Paper, URL: <http://www.kenorrinst.com/dwpaper.html>, 1996-1997.
- Parsaye, Kamran, et. al., *Intelligent Databases, Object-Oriented, Deductive Hypermedia Technologies*, John Wiley & Sons Inc., New York, NY, 1989.
- Sherman, Richard P., "Metadata: The Missing Link", *DBMS Online Magazine*, Vol. 10, No. 9, URL: <http://www.dbmsmag.com/9708d16.html>, August 1997.
- Strong, Diane M., et al., "10 Potholes in the Road to Information Quality", *IEEE Computer Magazine*, Vol. 10, No. 8, August 1997.

Turban, E., "Decision Support and Expert Systems", Management Support Systems, MacMillan Publishing, New York, NY, 1990.

Teorey, Toby J., Database Modeling & Design: The Fundamental Principles, Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1994.

## INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center .....2  
 8725 John J. Kingman Rd., STE 0944  
 Ft. Belvoir, Virginia, 22060-6218
  
2. Dudley Knox Library .....2  
 Naval Postgraduate School  
 411 Dyer Rd.  
 Monterey, California 93943-5101
  
3. Headquarters, U.S. Army Reserve Command .....2  
 ATTN: LTC Clevon  
 3800 North Camp Creek Parkway SW  
 Atlanta, Georgia 30331
  
4. Professor Daniel Dolk, Code SM/Dk.....3  
 Department of Systems Management  
 Naval Postgraduate School  
 555 Dyer Rd Bldg 330  
 Monterey, California 93943-5000
  
5. Professor George Thomas, Code SM/Te .....3  
 Department of Systems Management  
 Naval Postgraduate School  
 555 Dyer Rd Bldg 330  
 Monterey, California 93943-5000
  
6. LCDR Robert W. Dill .....3  
 386-C Bergin Drive  
 Monterey, California 93940